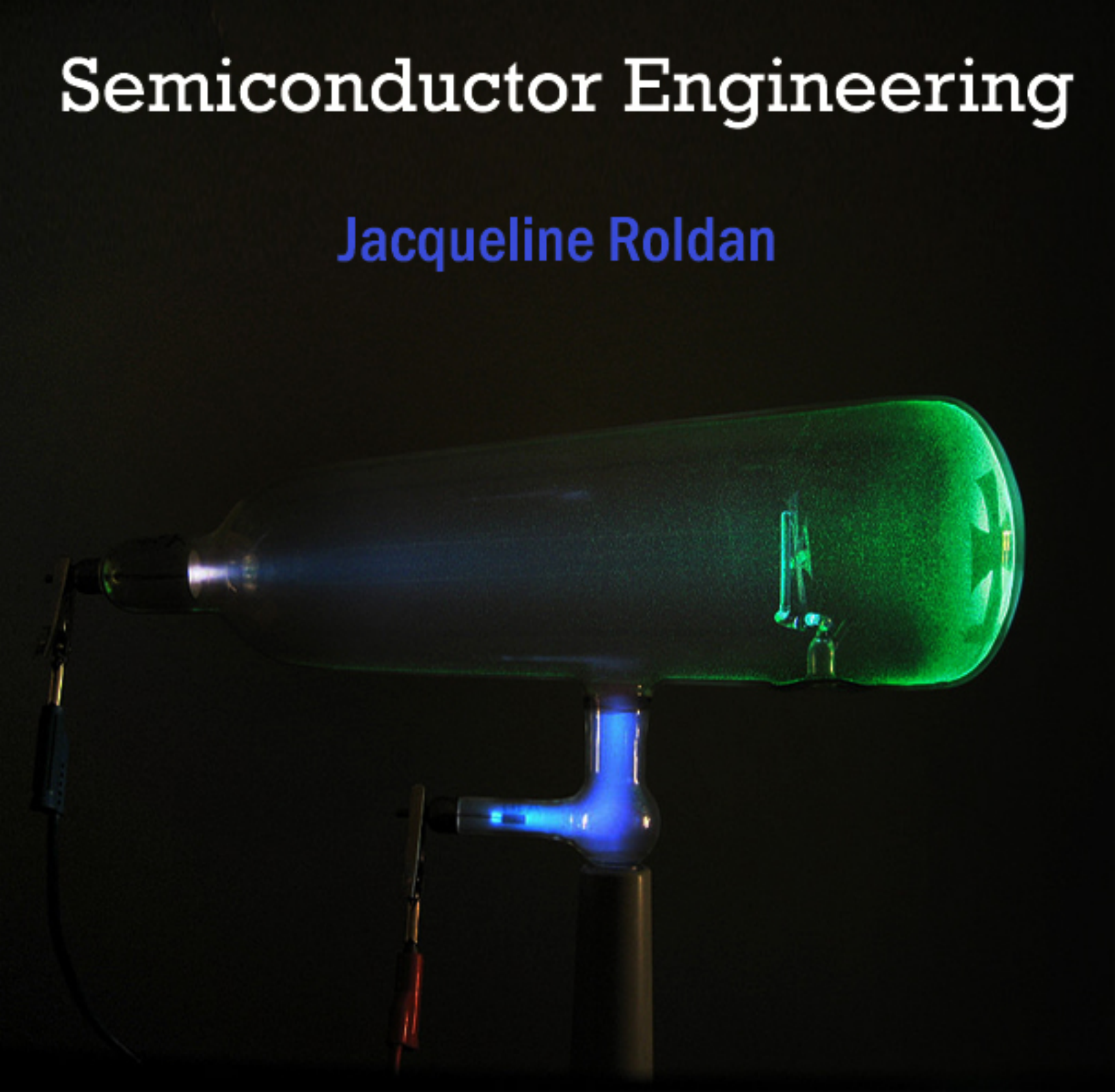
# Semiconductor Engineering

Jacqueline Roldan



First Edition, 2012

ISBN 978-81-323-4024-9

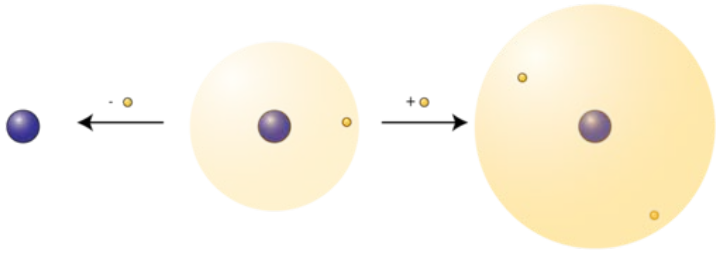


*Published by:*
**White Word Publications**
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

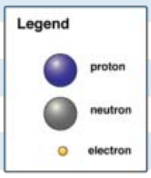
# Table of Contents

## Chapter-1

# Ion



| | | | |
|---|---|---|---|
| No. of protons | 1 | 1 | 1 |
| No. of electrons | 0 | 1 | 2 |
| Charge | +1 | 0 | -1 |
| Notation | $H^+$ | H | $H^-$ |
| Classification | cation | neutral (not an ion) | anion |




Hydrogen atom (center) contains a single proton and a single electron. Removal of the electron gives a cation (left), whereas addition of an electron gives an anion (right). The hydrogen anion, with its loosely held two-electron cloud, has a larger radius than the neutral atom, which in turn is much larger than the bare proton of the cation. Hydrogen forms the only cation that has no electrons, but even cations that (unlike hydrogen) still retain one or more electrons, are still smaller than the neutral atoms or molecules from which they are derived.

An **ion** is an atom or molecule in which the total number of electrons is not equal to the total number of protons, giving it a net positive or negative electrical charge. The name was given by physicist Michael Faraday for the substances that allow a current to pass ("go") between electrodes in a solution, when an electric field is applied. It is from Greek *ιον*, meaning "going."

An **anion** from the Greek word ἄνω (*άnō*), meaning "up", is an ion with more electrons than protons, giving it a net negative charge (since electrons are negatively charged and protons are positively charged). Conversely, a **cation** from the Greek word κατά (*katá*), meaning "down", is an ion with fewer electrons than protons, giving it a positive charge. Since the charge on a proton is equal in magnitude to the charge on an electron, the net

charge on an ion is equal to the number of protons in the ion minus the number of electrons.

An ion consisting of a single atom is an **atomic** or **monatomic ion**; if it consists of two or more atoms, it is a **molecular** or **polyatomic ion**.

## *General*

### History and Discovery

Etymologically the word *ion* is the Greek *ιον* (going), the present participle of *ιεναι*, *ienai*, "to go." This term was introduced by English physicist and chemist Michael Faraday in 1834 for the (then unknown) species that *goes* from one electrode to the other through an aqueous medium. Faraday did not know the nature of these species, but he knew that since metals dissolved into and entered solution at one electrode, and new metal came forth from solution at the other electrode, that some kind of substance moved through the solution in a current, conveying matter from one place to the other.

Faraday also introduced the words **anion** and **cation.** In Faraday's nomenclature, cations were named because they were attracted to the anode in a galvanic device and anions were named due to their attraction to the cathode.

### Characteristics

Ions in their gas-like state are highly reactive, and do not occur in large amounts on Earth, except in flames, lightning, electrical sparks, and other plasmas. These gas-like ions rapidly interact with ions of opposite charge to give neutral molecules or ionic salts. Ions are also produced in the liquid or solid state when salts interact with solvents (for example, water) to produce "solvated ions," which are more stable, for reasons involving a combination of energy and entropy changes as the ions move away from each other to interact with the liquid. These stabilized species are more commonly found in the environment at low temperatures. A common example is the ions present in seawater, which are derived from the dissolved salts there.

All ions are charged, which means that like all charged objects they are:

- attracted to opposite electric charges (positive to negative, and vice versa),
- repelled by like charges, and
- when moving, travel in trajectories that are deflected by a magnetic field.

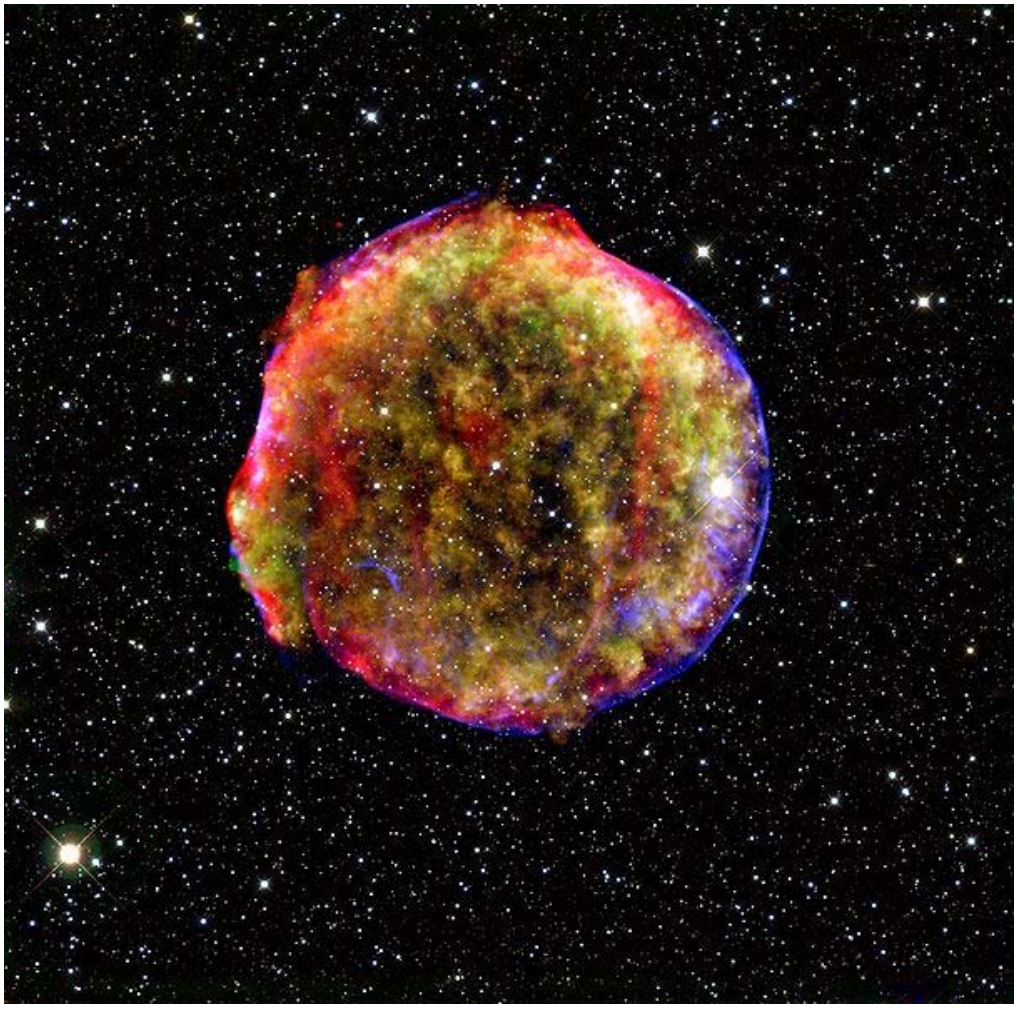
Electrons, due to their smaller mass and thus larger space-filling properties as matter waves, determine the size of atoms and molecules that possess any electrons at all. Thus, anions (negatively charged ions) are larger than the parent molecule or atom, as the excess electron(s) repel each other, and add to the physical size of the ion, because its size is determined by its electron cloud. Conversely, cations are generally smaller than the corresponding parent atom or molecule, for the same reason. One particular cation

(that of hydrogen) contains no electrons, and thus is *very much smaller* than the parent hydrogen atom.

## Natural Occurrences

Ions are ubiquitous in nature and are responsible for diverse phenomena from the luminescence of the Sun, and the existence of ionosphere on Earth. Atoms in their ionic state may have a different color from neutral atoms, and thus light absorption by metal ions gives the color of gemstones. In both inorganic and organic chemistry (including biochemistry), the interaction of water and ions is extremely important (an example is the energy that drives breakdown of ATP. The following sections describe contexts in which ions feature prominently and are arranged in decreasing physical length-scale, from the astronomical to the microscopic.

## Astronomical



The remnant of "Tycho's Supernova", a huge ball of expanding plasma. The outer shell shown in blue is X-ray emission by high-speed electrons.

A collection of non-aqueous gas-like ions, or even a gas containing a proportion of charged particles, is called a **plasma**. >99.9% of visible matter in the Universe may be in the form of plasmas. These include our Sun and other stars, the space between planets, as well as the space in between stars. Plasmas are often called the *fourth state of matter* because its properties are substantially different from solids, liquids, and gases. Astrophysical plasmas predominantly contain a mixture of electrons and protons (ionized hydrogen).

## Related Technology

Ions can be non-chemically prepared using various ion sources, usually involving high voltage or temperature. These are used in a multitude of devices such as mass spectrometers, optical emission spectrometers, particle accelerators, ion implanters and ion engines.

As reactive charged particles, they are also used in air purification by disrupting microbes, and in household items such as smoke detectors.

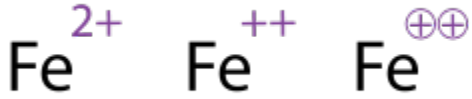
As signaling and metabolism in organisms are controlled by a precise ionic gradient across membranes, the disruption of this gradient contributes to cell death. This is a common mechanism exploited by natural and artificial biocides, including the ion channels gramicidin and amphotericin (a fungicide).

Inorganic dissolved ions are a component of total dissolved solids, an indicator of water quality in the world.

## *Chemistry*

### Notation

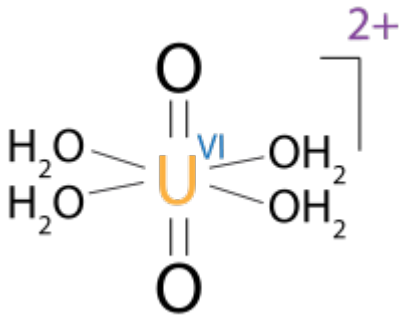
### Denoting the charged state




Equivalent notations for an iron atom (Fe) that lost two electrons.

When writing the chemical formula for an ion, its net charge is written in superscript immediately after the chemical structure for the molecule/atom. The net charge is written with the magnitude *before* the sign; that is, a doubly charged cation is indicated as **2+** instead of **+2**. Conventionally the magnitude of the charge is omitted for singly charged molecules/atoms; for example, the sodium cation is indicated as $Na^+$ and *not* $Na^{1+}$.

An alternative (and acceptable) way of showing a molecule/atom with multiple charges is by drawing out the signs multiple times; this is often seen with transition metals. Chemists sometimes circle the sign; this is merely ornamental and does not alter the chemical meaning. All three representations of $Fe^{2+}$ shown in the figure are thus equivalent.




**Mixed Roman numerals and charge notations for the uranyl ion**. The oxidation state of the metal is shown as superscripted Roman numerals, whereas the charge of the entire complex is shown by the angle symbol together with the magnitude and sign of the net charge.

Monatomic ions are sometimes also denoted with Roman numerals; for example, the $Fe^{2+}$ example seen above is occasionally referred to as Fe(II) or $Fe^{II}$. The Roman numeral designates the *formal oxidation state* of an element, whereas the superscripted numerals denotes the net charge. The two notations are therefore exchangeable for monatomic ions, but the Roman numerals *cannot* be applied to polyatomic ions. It is however possible to mix the notations for the individual metal center with a polyatomic complex, as shown by the uranyl ion example.

## Sub-classes

If an ion contains unpaired electrons, it is called a **radical** ion. Just like uncharged radicals, radical ions are very reactive. Polyatomic ions containing oxygen, such as carbonate and sulfate, are called **oxyanions**. Molecular ions that contain at least one carbon to hydrogen bond are called **organic ions**. If the charge in an organic ion is formally centered on a carbon, it is termed a **carbocation** (if positively charged) or **carboanion** (if negatively charged).

## Formation

## Formation of monatomic ions

Monatomic ions are formed by the addition of electrons to the valence shell of the atom, which is the outer-most electron shell in an atom, or the losing of electrons from this shell. The inner shells of an atom are filled with electrons that are tightly bound to the positively charged atomic nucleus, and so do not participate in this kind of chemical

interaction. The process of gaining or losing electrons from a neutral atom or molecule is called **ionization**.

Atoms can be ionized by bombardment with radiation, but the more usual process of ionization encountered in chemistry is the transfer of electrons between atoms or molecules. This transfer is usually driven by the attaining of stable ("closed shell") electronic configurations. Atoms will gain or lose electrons depending on which action takes the least energy.

For example, a sodium atom, Na, has a single electron in its valence shell, surrounding 2 stable, filled inner shells of 2 and 8 electrons. Since these filled shells are very stable, a sodium atom tends to lose its extra electron and attain this stable configuration, becoming a sodium **cation** in the process

$$Na \rightarrow Na^{+} + e-$$

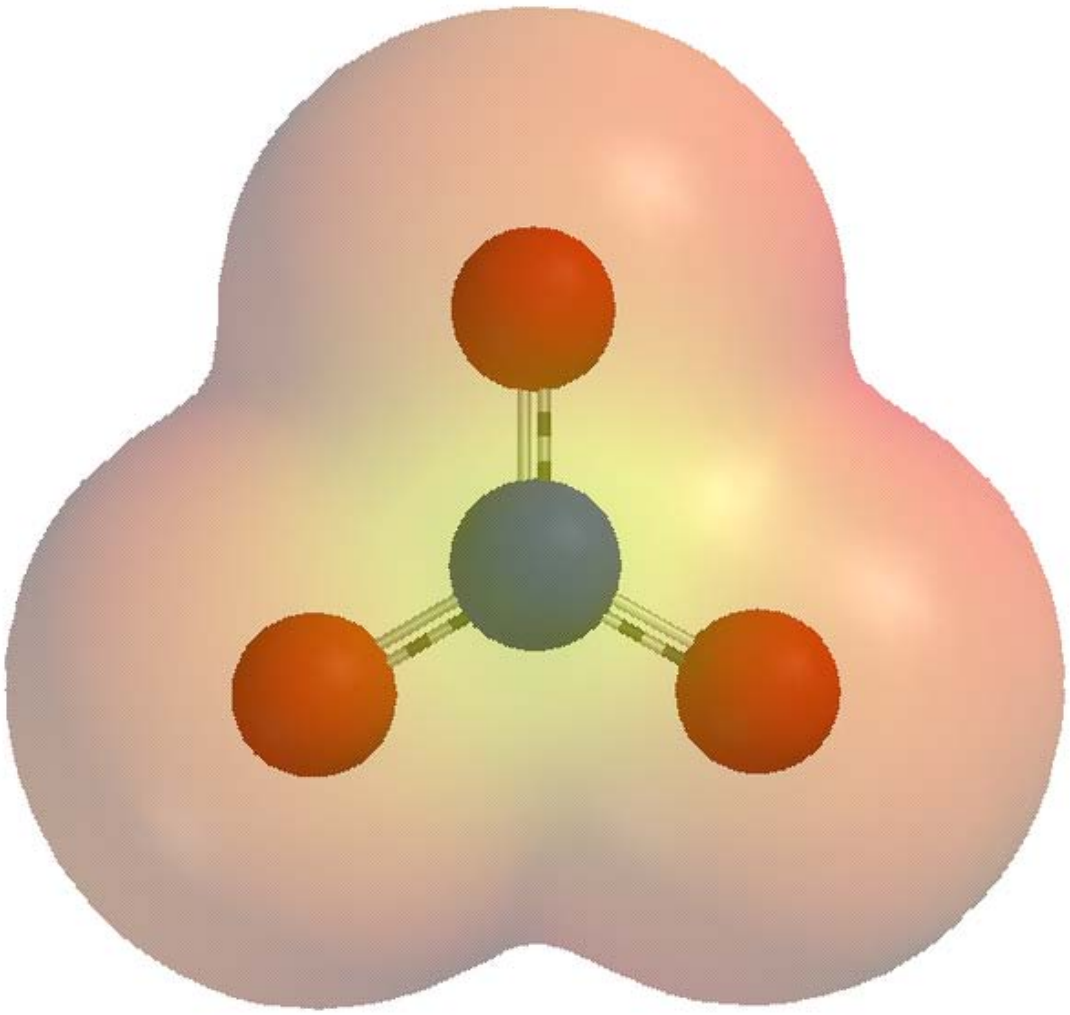
On the other hand, a chlorine atom, Cl, has 7 electrons in its valence shell, which is one short of the stable, filled shell with 8 electrons. Thus, a chlorine atom tends to *gain* an extra electron and attain a stable 8-electron configuration, becoming a chloride **anion** in the process:

$$Cl + e- \rightarrow Cl^{-}$$

This driving force is what causes sodium and chlorine to undergo a chemical reaction, where the "extra" electron is transferred from sodium to chlorine, forming sodium cations and chloride anions. Being oppositely charged, these cations and anions form ionic bonds and combine together to form sodium chloride, NaCl, more commonly known as rock salt.

$$Na^{+} + Cl^{-} \rightarrow NaCl$$

## Formation of polyatomic and molecular ions



An electrostatic potential map of the nitrate ion ($NO^{-}_{3}$). The 3-dimensional shell represents a single arbitrary isopotential.

Polyatomic and molecular ions are often formed by the gaining or losing of elemental ions such as $H^{+}$ in neutral molecules. For example, when ammonia, $NH_3$, accepts a proton, $H^{+}$, it forms the ammonium ion, $NH^{+}_{4}$. Ammonia and ammonium have the same number of electrons in essentially the same electronic configuration, but ammonium has an extra proton that gives it a net positive charge.

Ammonia can also lose an electron to gain a positive charge, forming the ion $\cdot NH^{+}_{3}$. However, this ion is unstable, because it has an incomplete valence shell around the nitrogen atom, making it a very reactive radical ion.

Due to the instability of radical ions, polyatomic and molecular ions are usually formed by gaining or losing elemental ions such as $H^+$, rather than gaining or losing electrons. This allows the molecule to preserve its stable electronic configuration while acquiring an electrical charge.

## Ionization potential

The energy required to detach an electron in its lowest energy state from an atom or molecule of a gas with less net electric charge is called the *ionization potential*, or *ionization energy*. The *n*th ionization energy of an atom is the energy required to detach its *n*th electron after the first $n - 1$ electrons have already been detached.

Each successive ionization energy is markedly greater than the last. Particularly great increases occur after any given block of atomic orbitals is exhausted of electrons. For this reason, ions tend to form in ways that leave them with full orbital blocks. For example, sodium has one *valence electron* in its outermost shell, so in ionized form it is commonly found with one lost electron, as $Na^+$. On the other side of the periodic table, chlorine has seven valence electrons, so in ionized form it is commonly found with one gained electron, as $Cl^-$. Caesium has the lowest measured ionization energy of all the elements and helium has the greatest. The ionization energy of metals is generally much lower than the ionization energy of nonmetals, which is why metals will generally lose electrons to form positively charged ions while nonmetals will generally gain electrons to form negatively charged ions.

## Ionic bonding

**Ionic bonding** is a kind of chemical bonding that arises from the mutual attraction of oppositely charged ions. Since ions of like charge repel each other, they do not usually exist on their own. Instead, many of them may form a crystal lattice, in which ions of opposite charge are bound to each other. The resulting compound is called an *ionic compound*, and is said to be held together by *ionic bonding*. In ionic compounds there arise characteristic distances between ion neighbors from which the spatial extension and the ionic radius of individual ions may be derived.

The most common type of ionic bonding is seen in compounds of metals and nonmetals (except noble gases, which rarely form chemical compounds). Metals are characterized by having a small number of electrons in excess of a stable, closed-shell electronic configuration. As such, they have the tendency to lose these extra electrons in order to attain a stable configuration. This property is known as *electropositivity*. Non-metals, on the other hand, are characterized by having an electron configuration just a few electrons short of a stable configuration. As such, they have the tendency to gain more electrons in order to achieve a stable configuration. This tendency is known as *electronegativity*. When a highly electropositive metal is combined with a highly electronegative nonmetal, the extra electrons from the metal atoms are transferred to the electron-deficient nonmetal atoms. This reaction produces metal cations and nonmetal anions, which are attracted to each other to form a *salt*.

## Common ions

Common **Cations**

| Common Name | Formula | Historic Name |
|---|---|---|
| ***Simple Cations*** | | |
| Aluminium | $Al^{3+}$ | |
| Calcium | $Ca^{2+}$ | |
| Copper(II) | $Cu^{2+}$ | cupric |
| Hydrogen | $H^{+}$ | |
| Iron(II) | $Fe^{2+}$ | ferrous |
| Iron(III) | $Fe^{3+}$ | ferric |
| Magnesium | $Mg^{2+}$ | |
| Mercury(II) | $Hg^{2+}$ | mercuric |
| Potassium | $K^{+}$ | kalic |
| Silver | $Ag^{+}$ | |
| Sodium | $Na^{+}$ | natric |
| ***Polyatomic Cations*** | | |
| Ammonium | $NH_4^{+}$ | |
| Oxonium | $H_3O^{+}$ | hydronium |
| Mercury(I) | $Hg_2^{2+}$ | mercurous |

Common **Anions**

| Formal Name | Formula | Alt. Name |
|---|---|---|
| ***Simple Anions*** | | |
| Chloride | $Cl^{-}$ | |
| Fluoride | $F^{-}$ | |
| Oxide | $O^{2-}$ | |
| ***Oxoanions*** | | |
| Carbonate | $CO_3^{2-}$ | |
| Hydrogen carbonate | $HCO_3^{-}$ | bicarbonate |
| Hydroxide | $OH^{-}$ | |
| Nitrate | $NO_3^{-}$ | |
| Phosphate | $PO_4^{3-}$ | |
| Sulfate | $SO_4^{2-}$ | |
| ***Anions from Organic Acids*** | | |
| Acetate | $CH_3COO^{-}$ | ethanoate |
| Formate | $HCOO^{-}$ | methanoate |
| Oxalate | $C_2O_4^{2-}$ | ethandioate |
| Cyanide | $CN^{-}$ | |

## Chapter-2

# Charge Carriers

In physics, a **charge carrier** denotes a free (mobile, unbound) particle carrying an electric charge, especially the particles that carry electric currents in electrical conductors. Examples are electrons and ions.

In metals, the charge carriers are electrons. One or two of the outer valence electrons from each atom is able to move about freely within the crystal lattice of the metal. This cloud of free electrons is referred to as a Fermi gas.

In ionic solutions, such as salt water, the charge carriers are the dissolved cations and anions. Similarly, cations and anions of the dissociated liquid serve as charge carriers in liquids and melted ionic solids.

In plasma, such as an electric arc, the electrons and cations of ionized gas and vaporized material of electrodes act as charge carriers. (The electrode vaporization occurs in vacuum too, but then the arc is not technically occurring in vacuum, but in low-pressure electrode vapors.)

In vacuum, in an electric arc or in vacuum tubes free electrons act as charge carriers.

In semiconductor physics, two types of charge carriers are recognized. One of them is electrons. In addition, it is convenient to treat the traveling vacancies in the valence-band electron population (holes) as the second type of charge carriers which carry the positive charge.

## *Free Carrier Concentration*

*Free carrier concentration* is the concentration of free carriers in a doped semiconductior. It is similar to the carrier concentration in a metal and for the purposes of calculating currents or drift velocities can be used in the same way. Free carriers are electrons (or holes) which have been introduced directly into the conduction (valence) band by doping and are not promoted thermally. For this reason electrons (holes) will not act as double carriers by leaving behind holes (electrons) in the other band.

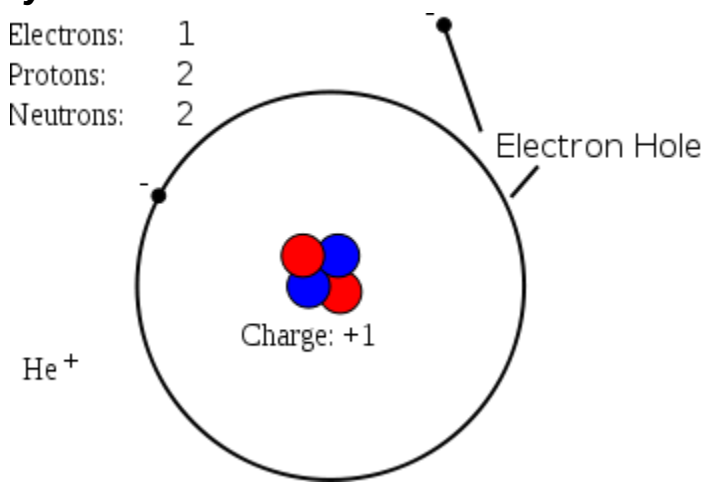
# Electron hole

An **electron hole** is the conceptual and mathematical opposite of an electron, useful in the study of physics, chemistry, and electrical engineering. The concept describes the lack of an electron at a position where one could exist in an atom or atomic lattice. It is different from the positron, which is the antimatter analogue of the electron.

The electron hole was introduced into calculations for the following two situations:

- If an electron is excited into a higher state it leaves a hole in its old state. This meaning is used in Auger electron spectroscopy (and other x-ray techniques), in computational chemistry, and to explain the low electron-electron scattering-rate in crystals (metals, semiconductors).
- In crystals, band structure calculations lead to an effective mass for the charge carriers, which can be negative. Inspired by the Hall effect, Newton's law is used to attach the negative sign onto the charge.

## *Solid state physics*




When an electron leaves a helium atom, it leaves an electron hole in its place. This causes the helium atom to become positively charged.

In solid state physics, an **electron hole** (usually referred to simply as a **hole**) is the absence of an electron from an otherwise full electron shell. A hole is essentially a way to conceptualise the interactions of the electrons within a nearly *full* system, which is *missing* just a few electrons. In some ways, the behaviour of a hole within a semiconductor crystal lattice is comparable to that of the bubble in an otherwise full bottle of water.

Hole conduction in a valence band can be explained by the following analogy. Imagine a row of people seated in an auditorium, where there are no spare chairs. Someone in the

middle of the row wants to leave, so he jumps over the back of the seat into an empty row, and walks out. The empty row is analogous to the conduction band, and the person walking out is analogous to a free electron.

Now imagine someone else comes along and wants to sit down. The empty row has a poor view; so he does not want to sit there. Instead, a person in the crowded row moves into the empty seat the first person left behind. The empty seat moves one spot closer to the edge and the person waiting to sit down. The next person follows, and the next, etcetera. One could say that the empty seat moves towards the edge of the row. Once the empty seat reaches the edge, the new person can sit down.

In the process everyone in the row has moved along. If those people were negatively charged (like electrons), this movement would constitute conduction. If the seats themselves were positively charged, then only the vacant seat would be positive. This is a very simple model of how hole conduction works.

In reality, due to the crystal structure properties, the hole is not localized to a single position as described in the previous example. Rather, the hole spans an area in the crystal lattice covering many hundreds of unit cells. This is equivalent to being unable to tell which broken bond corresponds to the "missing" electron.

Instead of analyzing the movement of an empty state in the valence band as the movement of billions of separate electrons, a single equivalent imaginary particle called a "hole" is considered. In an applied electric field, the electrons move in one direction, corresponding to the hole moving in the other. If a hole associates itself with a neutral atom, that atom loses an electron and becomes positive. Therefore the hole is taken to have positive charge of +e, precisely the opposite of the electron charge.

Coulomb's law allows the force on the "hole" due to an electric field to be calculated. An effective mass can then be derived which relates the (imaginary) force on the (imaginary) hole to the acceleration of that hole. In some semiconductors, such as silicon, the hole's effective mass is dependent on direction (anisotropic), however a value averaged over all directions can be used for some macroscopic calculations.

In most semiconductors, the effective mass of a hole is much larger than that of an electron. This results in lower mobility for holes under the influence of an electric field and this may slow down the speed of the electronic device made of that semiconductor. This is one major reason for adopting electrons as the primary charge carriers, whenever possible in semiconductor devices instead of holes.

## *Holes in quantum chemistry*

An alternate meaning for the term **electron hole** is used in computational chemistry. In coupled cluster methods, the ground (or lowest energy) state of a molecule is interpreted as the "vacuum state"—conceptually, in this state there are no electrons. In this scheme, the absence of an electron from a normally-filled state is called a "hole" and is treated as

a particle, and the presence of an electron in a normally-empty state is simply called an "electron". This terminology is almost identical to that used in solid-state physics.

# Non-radiative recombination

**Non-radiative recombination** is a process in phosphors and semiconductors, whereby charge carriers recombine without releasing photons. A phonon is released instead.

Non-radiative recombination in optoelectronics and phosphors is an unwanted process, lowering the light generation efficiency and increasing heat losses.

## *Shockley–Read–Hall (SRH) process*

The electron in transition between bands passes through a new energy state created within the band gap by an impurity in the crystal lattice. The impurity state can absorb differences in momentum between the carriers, and so this process is the dominant generation and recombination process in silicon and other indirect bandgap materials. It can also dominate in direct bandgap materials under conditions of very low carrier densities (very low level injection). The energy is exchanged in the form of lattice vibration, or a phonon exchanging thermal energy with the material.

Various impurities and dislocations create energy levels within the band gap corresponding to neither donor nor acceptor levels, forming deep-level traps. Non-radiative recombination occurs primarily at such sites.

# Charge carrier density

The **charge carrier density** denotes the number of charge carriers per volume. It is measured in $m^{-3}$. As any density it can depend on position. It should not be confused with the charge density which is the number of charges per volume at a given energy.

The carrier density is obtained by integrating the charge density over the energy that the charges are allowed to have.

The charge carrier density is a particle density, so integrating it over a volume $V$ gives the number of charge carriers $N$ in that volume

$$N = \int_V n(\mathbf{r})\,\mathrm{d}V$$

.

where

$n(\mathbf{r})$ is the position-dependent charge carrier density.

If the density does not depend on position and is instead equal to a constant $n_0$ this equation simplifies to

$$N = V \cdot n_0.$$

The charge carrier densities enters equations concerning the electrical conductivity and related phenomena like the thermal conductivity.

# Charge carriers in semiconductors

There are two recognized types of **charge carriers in semiconductors**. One of them is electrons, which carry negative electric charge. In addition, it is convenient to treat the traveling vacancies in the valence-band electron population (holes) as the second type of charge carriers, which carry a positive charge equal in magnitude to that of an electron.

## *Carrier generation and recombination*

When an electron meets with a hole, they recombine and these free carriers effectively vanish. The energy released can be either thermal, heating up the semiconductor (*thermal recombination*, one of the sources of waste heat in semiconductors), or released as photons (*optical recombination*, used in LEDs and semiconductor lasers).

### Majority and minority carriers

The more abundant charge carriers are called **majority carriers**. In n-type semiconductors they are electrons, while in p-type semiconductors they are holes. The less abundant charge carriers are called **minority carriers**; in n-type semiconductors they are holes, while in p-type semiconductors they are electrons.

In an intrinsic semiconductor the concentrations of both types of carriers are ideally equal.

Minority carriers play an important role in bipolar transistors and solar cells. However, their role in field-effect transistors (FETs) is a bit more complex: for example, a MOSFET has both p-type and n-type regions. The transistor action involves the majority carriers of the source and drain regions, but these carriers traverse the body of the opposite type, where they are minority carriers. However, the traversing carriers hugely outnumber their opposite type in the transfer region (in fact, the opposite type carriers are removed by an applied electric field that creates an inversion layer), so conventionally the

source and drain designation for the carriers is adopted, and FETs are called "majority carrier" devices.

# Saturation velocity

**Saturation velocity** is the maximum velocity a charge carrier in a semiconductor, generally an electron, attains in the presence of very high electric fields. Charge carriers normally move at an average drift speed proportional to the electric field strength they experience temporally. The proportionality constant is known as mobility of the carrier, which is a material property. A good conductor would have a high mobility value for its charge carrier, which means higher velocity, and consequently higher current values for a given electric field strength. There is a limit though to this process and at some high field value, a charge carrier can not move any faster, having reached its saturation velocity, due to mechanisms that eventually limit the movement of the carriers in the material.

Saturation velocity is a very important parameter in the design of semiconductor devices, especially field effect transistors, which are basic building blocks of almost all modern integrated circuits. Typical values of saturation velocity may vary greatly for different materials, for example for Si it is in the order of $1\times10^7$ cm/s, for GaAs $1.2\times10^7$ cm/s, while for 6H-SiC, it is near $2\times10^7$ cm/s. Typical electric field strengths at which carrier velocity saturates is usually on the order of 10-100 kV/cm. Both saturation field and the saturation velocity of a semiconductor material are typically strong function of impurities, crystal defects and temperature.

For extremely small scale devices, where the high-field regions may be comparable or smaller than the average mean free path of the charge carrier, one can observe velocity overshoot, or hot electron effects which has become more important as the transistor geometries continually decrease to enable design of faster, larger and more dense integrated circuits. The regime where the two terminals between which the elctron moves is much smaller than the mean free path, is sometimes referred as ballistic transport. There has been numerous attempts in the past to build transistors based on this principle without much success. Nevertheless, developing field of nanotechnology, and new materials such as Carbon nanotubes and graphene, offers new hope.

Though in a semiconductor such as Si saturation velocity of a carrier is same as the peak velocity of the carrier, for some other materials with more complex energy band structures, this is not true. In GaAs or InP for example the carrier drift velocity reaches to a maximum as a function of field and then it begins to actually decrease as the electric field applied is increased further. Carriers which have gained enough energy are kicked up to a different conduction band which presents a lower drift velocity and eventually a lower saturation velocity in these materials. This results in an overall decrease of current for higher voltage until all electrons are in the "slow" band and this is the principle behind operation of a Gunn diode, which can display negative differential resistivity. Due

to the transfer of electrons to a different conduction band involved, such devices, usually single terminal, are referred to as transferred electron devices, or TEDs.

# Spontaneous emission

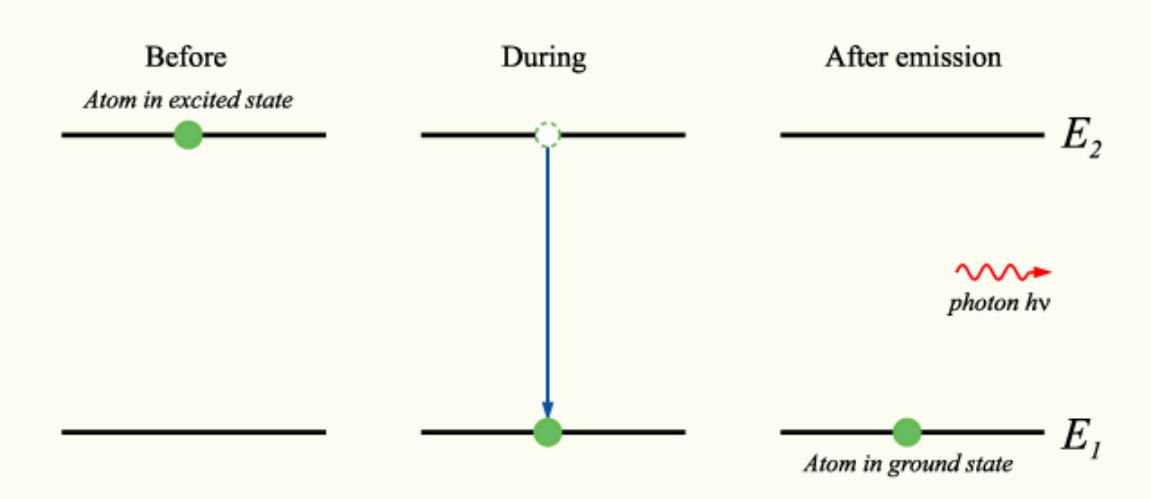
**Spontaneous emission** is the process by which a light source such as an atom, molecule, nanocrystal or nucleus in an excited state undergoes a transition to a state with a lower energy, e.g., the ground state and emits a photon. Spontaneous emission of light or luminescence is a fundamental process that plays an essential role in many phenomena in nature and forms the basis of many applications, such as fluorescent tubes, older television screens (cathode ray tubes), plasma display panels, lasers (for startup - normal continuous operation works by stimulated emission instead) and light emitting diodes.

## *Introduction*

If a light source ('the atom') is in the excited state with energy $E_2$, it may spontaneously decay to a lower lying level (e.g., the ground state) with energy $E_1$, releasing the difference in energy between the two states as a photon. The photon will have angular frequency ω and energy $\hbar\omega$(= $h\nu$, where $h$ is the Planck constant and ν is the frequency):

$$E_2 - E_1 = \hbar\omega,$$

where $\hbar$is the reduced Planck constant. The phase of the photon in spontaneous emission is random as is the direction the photon propagates in. This is not true for stimulated emission. An energy level diagram illustrating the process of spontaneous emission is shown below:



If the number of light sources in the excited state is given by $N$, the rate at which $N$ decays is:

$$\frac{\partial N}{\partial t} = -A_{21} N,$$

where $A_{21}$ is the rate of spontaneous emission. In the rate-equation $A_{21}$ is a proportionality constant for this particular transition in this particular light source. The constant is referred to as the *Einstein A coefficient*, and has units $s^{-1}$. The above equation can be solved to give:

$$N(t) = N(0)e^{-A_{21}t} = N(0)e^{-\Gamma_{rad}t},$$

where $N(0)$ is the initial number of light sources in the excited state, $t$ is the time and $\Gamma_{rad}$ is the radiative decay rate of the transition. The number of excited states $N$ thus decays exponentially with time, similar to radioactive decay. After one lifetime, the number of excited states decays to 36.8% of its original value ($\frac{1}{e}$-time). The radiative decay rate $\Gamma_{rad}$ is inversely proportional to the lifetime $\tau_{12}$:

$$A_{21} = \Gamma_{12} = \frac{1}{\tau_{21}}.$$

## Theory

Spontaneous transitions was not explainable within the framework of the old quantum theory, that is a theory in which the atomic levels are quantized, but the electromagnetic field is not. In fact, using the machinery of the usually called "first-quantized" quantum mechanics and one computes the probability of spontaneous transitions from one stationary state to another, one finds that it is zero. In order to explain spontaneous transitions, quantum mechanics must be extended to a "second-quantized" theory, wherein the electromagnetic field is quantized at every point in space. Such a theory is known as a quantum field theory; the quantum field theory of electrons and electromagnetic fields is known as quantum electrodynamics.

In quantum electrodynamics (or QED), the electromagnetic field has a ground state, the vacuum state, which can mix with the excited stationary states of the atom. As a result of this interaction, the "stationary state" of the atom is no longer a true eigenstate of the combined system of the atom plus electromagnetic field. In particular, the electron transition from the excited state to the electronic ground state mixes with the transition of the electromagnetic field from the ground state to an excited state, a field state with one photon in it. Spontaneous emission in free space depends upon vacuum fluctuations to get started.

Although there is only one electronic transition from the excited state to ground state, there are many ways in which the electromagnetic field may go from the ground state to a one-photon state. That is, the electromagnetic field has infinitely more degrees of freedom, corresponding to the different directions in which the photon can be emitted. Equivalently, one might say that the phase space offered by the electromagnetic field is infinitely larger than that offered by the atom. Since one must consider probabilities that occupy all of phase space equally, the combined system of atom plus electromagnetic field must undergo a transition from electronic excitation to a photonic excitation; the atom must decay by spontaneous emission. The time the light source remains in the excited state thus depends on the light source itself as well as its environment. Imagine trying to hold a pencil upright on the end of your finger. It will stay there if your hand is perfectly stable and nothing perturbs the equilibrium. But the slightest perturbation will make the pencil fall into a more stable equilibrium position. Similarly, vacuum fluctuations cause an excited atom to fall into its ground state.

In spectroscopy one can frequently find that atoms or molecules in the excited states dissipate their energy in the absence of any external source of photons. This is not spontaneous emission, but is actually nonradiative relaxation of the atoms or molecules caused by the fluctuation of the surrounding molecules present inside the bulk.

## *Rate of spontaneous emission*

The rate of spontaneous emission (i.e., the radiative rate) can be described by Fermi's golden rule. The rate of emission depends on two factors: an 'atomic part', which describes the internal structure of the light source and a 'field part', which describes the density of electromagnetic modes of the environment. The atomic part describes the strength of a transition between two states in terms of transition moments. In a homogeneous medium, such as free space, the rate of spontaneous emission in the dipole approximation is given by:

$$\Gamma_{rad}(\omega) = \frac{\omega^3 n |\mu_{12}|^2}{3\pi\varepsilon_0\hbar {c_0}^3}$$

where ω is the emission frequency, $n$ is the index of refraction, $\mu_{12}$ is the transition dipole moment, $\varepsilon_0$ is the vacuum permittivity, $\hbar$ is the reduced Planck constant and $c_0$ is the vacuum speed of light. (This approximation breaks down in the case of inner shell electrons in high-Z atoms.) Clearly, the rate of spontaneous emission in free space increases with $\omega^3$. In contrast with atoms, which have a discrete emission spectrum, quantum dots can be tuned continuously by changing their size. This property has been used to check the $\omega^3$-frequency dependence of the spontaneous emission rate as described by Fermi's golden rule.

## *Radiative and nonradiative decay: the quantum efficiency*

In the rate-equation above, it is assumed that decay of the number of excited states $N$ only occurs under emission of light. In this case one speaks of full radiative decay and this

means that the quantum efficiency is 100%. Besides radiative decay, which occurs under the emission of light, there is a second decay mechanism; nonradiative decay. To determine the total decay rate $\Gamma_{tot}$, radiative and nonradiative rates should be summed:

$$\Gamma_{tot} = \Gamma_{rad} + \Gamma_{nrad}$$

where $\Gamma_{tot}$ is the total decay rate, $\Gamma_{rad}$ is the radiative decay rate and $\Gamma_{nrad}$ the nonradiative decay rate. The quantum efficiency (QE) is defined as the fraction of emission processes in which emission of light is involved:

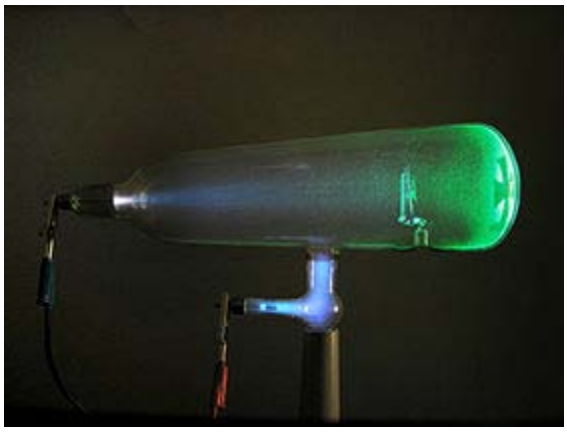
$$QE = \frac{\Gamma_{rad}}{\Gamma_{nrad} + \Gamma_{rad}}.$$

In nonradiative relaxation, the energy is released as phonons, more commonly known as heat. Nonradiative relaxation occurs when the energy difference between the levels is very small, and these typically occur on a much faster time scale than radiative transitions. For many materials (for instance, semiconductors), electrons move quickly from a high energy level to a meta-stable level via small nonradiative transitions and then make the final move down to the bottom level via an optical or radiative transition. This final transition is the transition over the bandgap in semiconductors. Large nonradiative transitions do not occur frequently because the crystal structure generally can not support large vibrations without destroying bonds (which generally doesn't happen for relaxation). Meta-stable states form a very important feature that is exploited in the construction of lasers. Specifically, since electrons decay slowly from them, they can be piled up in this state without too much loss and then stimulated emission can be used to boost an optical signal.

## Chapter-3

# Electron

***Electron***



Experiments with a Crookes tube first demonstrated the particle nature of electrons. In this illustration, the profile of the cross-shaped target is projected against the tube face at right by a beam of electrons.

| | |
|---|---|
| **Composition:** | Elementary particle |
| **Particle statistics:** | Fermionic |
| **Group:** | Lepton |
| **Generation:** | First |
| **Interaction:** | Gravity, Electromagnetic, Weak |
| **Symbol(s):** | $e^-$, $\beta^-$ |
| **Antiparticle:** | Positron (also called antielectron) |
| **Theorized:** | Richard Laming (1838–1851), G. Johnstone Stoney (1874) and others. |
| **Discovered:** | J. J. Thomson (1897) |
| **Mass:** | $9.10938215(45)\times10^{-31}$ kg |

| | |
|---|---|
| | $5.4857990943(23)\times10^{-4}$ u |
| | $[1{,}822.88850204(77)]^{-1}$ u |
| | 0.510998910(13) MeV/$c^2$ |
| **Electric charge:** | −1 e |
| | $-1.602176487(40)\times10^{-19}$ C |
| | $-4.803\times10^{-10}$ esu |
| **Magnetic moment:** | $-1.00115965218111\ \mu_B$ |
| **Spin:** | $\frac{1}{2}$ |

The **electron** is a subatomic particle carrying a negative electric charge. It has no known components or substructure. Therefore, the electron is generally believed to be an elementary particle. An electron has a mass that is approximately 1/1836 that of the proton The intrinsic angular momentum (spin) of the electron is a half-integer value in units of $\hbar$, which means that it is a fermion. The antiparticle of the electron is called the positron. The positron is identical to the electron except that it carries electrical and other charges of the opposite sign. When an electron collides with a positron, both particles may either scatter off each other or be totally annihilated, producing a pair (or more) of gamma ray photons. Electrons, which belong to the first generation of the lepton particle family, participate in gravitational, electromagnetic and weak interactions. Electrons, like all matter, have quantum mechanical properties of both particles and waves, so they can collide with other particles and be diffracted like light. However, this duality is best demonstrated in experiments with electrons, due to their tiny mass. Since an electron is a fermion, no two electrons can occupy the same quantum state, in accordance with the Pauli exclusion principle.

The concept of an indivisible amount of electric charge was theorized to explain the chemical properties of atoms, beginning in 1838 by British natural philosopher Richard Laming; the name *electron* was introduced for this charge in 1894 by Irish physicist George Johnstone Stoney. The electron was identified as a particle in 1897 by J. J. Thomson and his team of British physicists.

In many physical phenomena, such as electricity, magnetism, and thermal conductivity, electrons play an essential role. An electron in motion relative to an observer generates a magnetic field, and will be deflected by external magnetic fields. When an electron is accelerated, it can absorb or radiate energy in the form of photons. Electrons, together with atomic nuclei made of protons and neutrons, make up atoms. However, electrons contribute less than 0.06% to an atom's total mass. The attractive Coulomb force between an electron and a proton causes electrons to be bound into atoms. The exchange or sharing of the electrons between two or more atoms is the main cause of chemical bonding.

According to theory, most electrons in the universe were created in the big bang, but they may also be created through beta decay of radioactive isotopes and in high-energy collisions, for instance when cosmic rays enter the atmosphere. Electrons may be destroyed through annihilation with positrons, and may be absorbed during nucleosynthesis in stars. Laboratory instruments are capable of containing and observing individual electrons as well as electron plasma, whereas dedicated telescopes can detect electron plasma in outer space. Electrons have many applications, including welding, cathode ray tubes, electron microscopes, radiation therapy, lasers and particle accelerators.

## History

The ancient Greeks noticed that amber attracted small objects when rubbed with fur. Apart from lightning, this phenomenon is humanity's earliest recorded experience with electricity. In his 1600 treatise *De Magnete*, the English scientist William Gilbert coined the New Latin term *electricus*, to refer to this property of attracting small objects after being rubbed. Both *electric* and *electricity* are derived from the Latin *ēlectrum* (also the root of the alloy of the same name), which came from the Greek word ήλεκτρον (*ēlektron*) for amber.

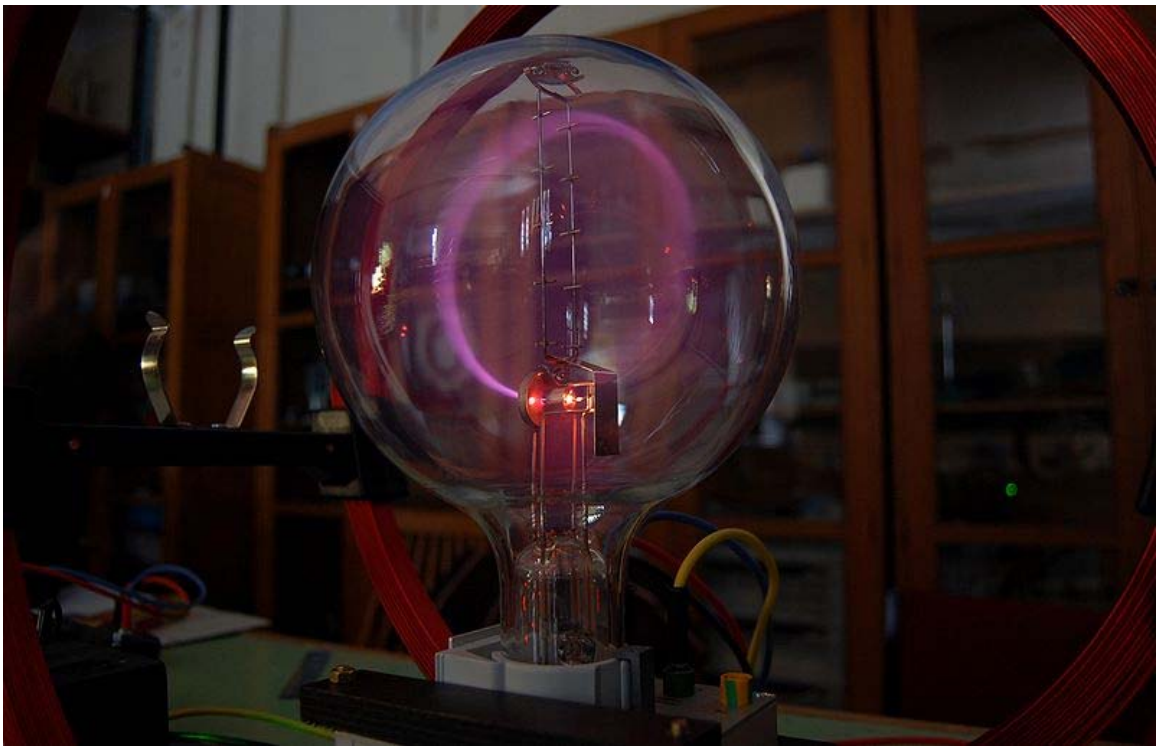
In 1737 C. F. du Fay and Hawksbee independently discovered what they believed to be two kinds of frictional electricity; one generated from rubbing glass, the other from rubbing resin. From this, Du Fay theorized that electricity consists of two electrical fluids, "vitreous" and "resinous", that are separated by friction and that neutralize each other when combined. A decade later Benjamin Franklin proposed that electricity was not from different types of electrical fluid, but the same electrical fluid under different pressures. He gave them the modern charge nomenclature of positive and negative respectively. Franklin thought that the charge carrier was positive.

Between 1838 and 1851, British natural philosopher Richard Laming developed the idea that an atom is composed of a core of matter surrounded by subatomic particles that had unit electric charges. Beginning in 1846, German physicist William Weber theorized that electricity was composed of positively and negatively charged fluids, and their interaction was governed by the inverse square law. After studying the phenomenon of electrolysis in 1874, Irish physicist George Johnstone Stoney suggested that there existed a "single definite quantity of electricity", the charge of a monovalent ion. He was able to estimate the value of this elementary charge *e* by means of Faraday's laws of electrolysis. However, Stoney believed these charges were permanently attached to atoms and could not be removed. In 1881, German physicist Hermann von Helmholtz argued that both positive and negative charges were divided into elementary parts, each of which "behaves like atoms of electricity".

In 1894, Stoney coined the term *electron* to describe these elementary charges, saying, "... an estimate was made of the actual amount of this most remarkable fundamental unit of electricity, for which I have since ventured to suggest the name *electron*". The word

*electron* is a combination of the word *electric* and the suffix *-on*, with the latter now used to designate a subatomic particle, such as a proton or neutron.

## Discovery



A beam of electrons deflected in a circle by a magnetic field

The German physicist Johann Wilhelm Hittorf undertook the study of electrical conductivity in rarefied gases. In 1869, he discovered a glow emitted from the cathode that increased in size with decrease in gas pressure. In 1876, the German physicist Eugen Goldstein showed that the rays from this glow cast a shadow, and he dubbed the rays cathode rays. During the 1870s, the English chemist and physicist Sir William Crookes developed the first cathode ray tube to have a high vacuum inside. He then showed that the luminescence rays appearing within the tube carried energy and moved from the cathode to the anode. Furthermore, by applying a magnetic field, he was able to deflect the rays, thereby demonstrating that the beam behaved as though it were negatively charged. In 1879, he proposed that these properties could be explained by what he termed 'radiant matter'. He suggested that this was a fourth state of matter, consisting of negatively charged molecules that were being projected with high velocity from the cathode.

The German-born British physicist Arthur Schuster expanded upon Crookes' experiments by placing metal plates in parallel to the cathode rays and applying an electric potential between the plates. The field deflected the rays toward the positively charged plate, providing further evidence that the rays carried negative charge. By measuring the

amount of deflection for a given level of current, in 1890 Schuster was able to estimate the charge-to-mass ratio of the ray components. However, this produced a value that was more than a thousand times greater than what was expected, so little credence was given to his calculations at the time.

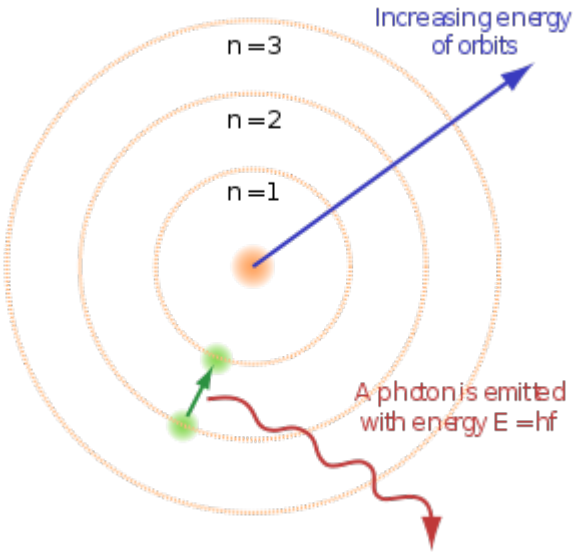
In 1896, the British physicist J. J. Thomson, with his colleagues John S. Townsend and H. A. Wilson, performed experiments indicating that cathode rays really were unique particles, rather than waves, atoms or molecules as was believed earlier. Thomson made good estimates of both the charge *e* and the mass *m*, finding that cathode ray particles, which he called "corpuscles," had perhaps one thousandth of the mass of the least massive ion known: hydrogen. He showed that their charge to mass ratio, *e*/*m*, was independent of cathode material. He further showed that the negatively charged particles produced by radioactive materials, by heated materials and by illuminated materials were universal. The name electron was again proposed for these particles by the Irish physicist George F. Fitzgerald, and the name has since gained universal acceptance.

While studying naturally fluorescing minerals in 1896, the French physicist Henri Becquerel discovered that they emitted radiation without any exposure to an external energy source. These radioactive materials became the subject of much interest by scientists, including the New Zealand physicist Ernest Rutherford who discovered they emitted particles. He designated these particles alpha and beta, on the basis of their ability to penetrate matter. In 1900, Becquerel showed that the beta rays emitted by radium could be deflected by an electric field, and that their mass-to-charge ratio was the same as for cathode rays. This evidence strengthened the view that electrons existed as components of atoms.

The electron's charge was more carefully measured by the American physicist Robert Millikan in his oil-drop experiment of 1909, the results of which he published in 1911. This experiment used an electric field to prevent a charged droplet of oil from falling as a result of gravity. This device could measure the electric charge from as few as 1–150 ions with an error margin of less than 0.3%. Comparable experiments had been done earlier by Thomson's team, using clouds of charged water droplets generated by electrolysis, and in 1911 by Abram Ioffe, who independently obtained the same result as Millikan using charged microparticles of metals, then published his results in 1913. However, oil drops were more stable than water drops because of their slower evaporation rate, and thus more suited to precise experimentation over longer periods of time.

Around the beginning of the twentieth century, it was found that under certain conditions a fast moving charged particle caused a condensation of supersaturated water vapor along its path. In 1911, Charles Wilson used this principle to devise his cloud chamber, allowing the tracks of charged particles, such as fast-moving electrons, to be photographed.

## Atomic theory




The Bohr model of the atom, showing states of electron with energy quantized by the number n. An electron dropping to a lower orbit emits a photon equal to the energy difference between the orbits.

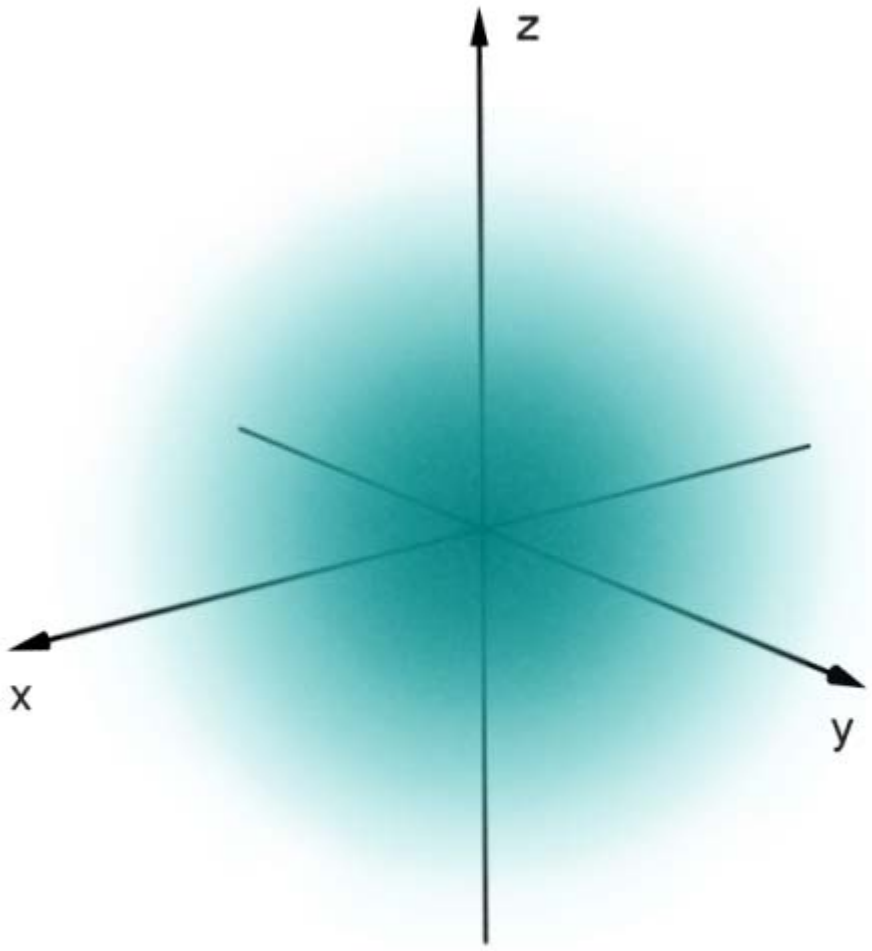
By 1914, experiments by physicists Ernest Rutherford, Henry Moseley, James Franck and Gustav Hertz had largely established the structure of an atom as a dense nucleus of positive charge surrounded by lower-mass electrons. In 1913, Danish physicist Niels Bohr postulated that electrons resided in quantized energy states, with the energy determined by the angular momentum of the electron's orbits about the nucleus. The electrons could move between these states, or orbits, by the emission or absorption of photons at specific frequencies. By means of these quantized orbits, he accurately explained the spectral lines of the hydrogen atom. However, Bohr's model failed to account for the relative intensities of the spectral lines and it was unsuccessful in explaining the spectra of more complex atoms.

Chemical bonds between atoms were explained by Gilbert Newton Lewis, who in 1916 proposed that a covalent bond between two atoms is maintained by a pair of electrons shared between them. Later, in 1923, Walter Heitler and Fritz London gave the full explanation of the electron-pair formation and chemical bonding in terms of quantum mechanics. In 1919, the American chemist Irving Langmuir elaborated on the Lewis' static model of the atom and suggested that all electrons were distributed in successive "concentric (nearly) spherical shells, all of equal thickness". The shells were, in turn, divided by him in a number of cells each containing one pair of electrons. With this model Langmuir was able to qualitatively explain the chemical properties of all elements in the periodic table, which were known to largely repeat themselves according to the periodic law.

In 1924, Austrian physicist Wolfgang Pauli observed that the shell-like structure of the atom could be explained by a set of four parameters that defined every quantum energy state, as long as each state was inhabited by no more than a single electron. (This prohibition against more than one electron occupying the same quantum energy state became known as the Pauli exclusion principle.) The physical mechanism to explain the fourth parameter, which had two distinct possible values, was provided by the Dutch physicists Abraham Goudsmith and George Uhlenbeck when they suggested that an electron, in addition to the angular momentum of its orbit, could possess an intrinsic angular momentum. This property became known as spin, and explained the previously mysterious splitting of spectral lines observed with a high-resolution spectrograph; this phenomenon is known as fine structure splitting.

## Quantum mechanics

In his 1924 dissertation *Recherches sur la théorie des quanta* (Research on Quantum Theory), French physicist Louis de Broglie hypothesized that all matter possesses a De Broglie wave similar to light. That is, under the appropriate conditions, electrons and other matter would show properties of either particles or waves. The corpuscular properties of a particle are demonstrated when it is shown to have a localized position in space along its trajectory at any given moment. Wave-like nature is observed, for example, when a beam of light is passed through parallel slits and creates interference patterns. In 1927, the interference effect was demonstrated with a beam of electrons by English physicist George Paget Thomson with a thin metal film and by American physicists Clinton Davisson and Lester Germer using a crystal of nickel.




Orbital s ($\ell = 0$, $m_\ell = 0$)

In quantum mechanics, the behavior of an electron in an atom is described by an orbital, which is a probability distribution rather than an orbit. In the figure, the shading indicates the relative probability to "find" the electron, having the energy corresponding to the given quantum numbers, at that point.

The success of de Broglie's prediction led to the publication, by Erwin Schrödinger in 1926, of the Schrödinger equation that successfully describes how electron waves propagated. Rather than yielding a solution that determines the location of an electron over time, this wave equation can be used to predict the probability of finding an electron near a position. This approach was later called quantum mechanics, which provided an extremely close derivation to the energy states of an electron in a hydrogen atom. Once spin and the interaction between multiple electrons were considered, quantum mechanics allowed the configuration of electrons in atoms with higher atomic numbers than hydrogen to be successfully predicted.

In 1928, building on Wolfgang Pauli's work, Paul Dirac produced a model of the electron - the Dirac equation, consistent with relativity theory, by applying relativistic and symmetry considerations to the hamiltonian formulation of the quantum mechanics of the electro-magnetic field. In order to resolve some problems within his relativistic equation, in 1930 Dirac developed a model of the vacuum as an infinite sea of particles having negative energy, which was dubbed the Dirac sea. This led him to predict the existence of a positron, the antimatter counterpart of the electron. This particle was discovered in 1932 by Carl D. Anderson, who proposed calling standard electrons *negatrons*, and using *electron* as a generic term to describe both the positively and negatively charged variants. This usage of the term 'negatron' is still occasionally encountered today, and it may be shortened to 'negaton'.

In 1947 Willis Lamb, working in collaboration with graduate student Robert Rutherford, found that certain quantum states of hydrogen atom, which should have the same energy, were shifted in relation to each other, the difference being the Lamb shift. About the same time, Polykarp Kusch, working with Henry M. Foley, discovered the magnetic moment of the electron is slightly larger than predicted by Dirac's theory. This small difference was later called anomalous magnetic dipole moment of the electron. To resolve these issues, a refined theory called quantum electrodynamics was developed by Sin-Itiro Tomonaga, Julian Schwinger and Richard P. Feynman in the late 1940s.

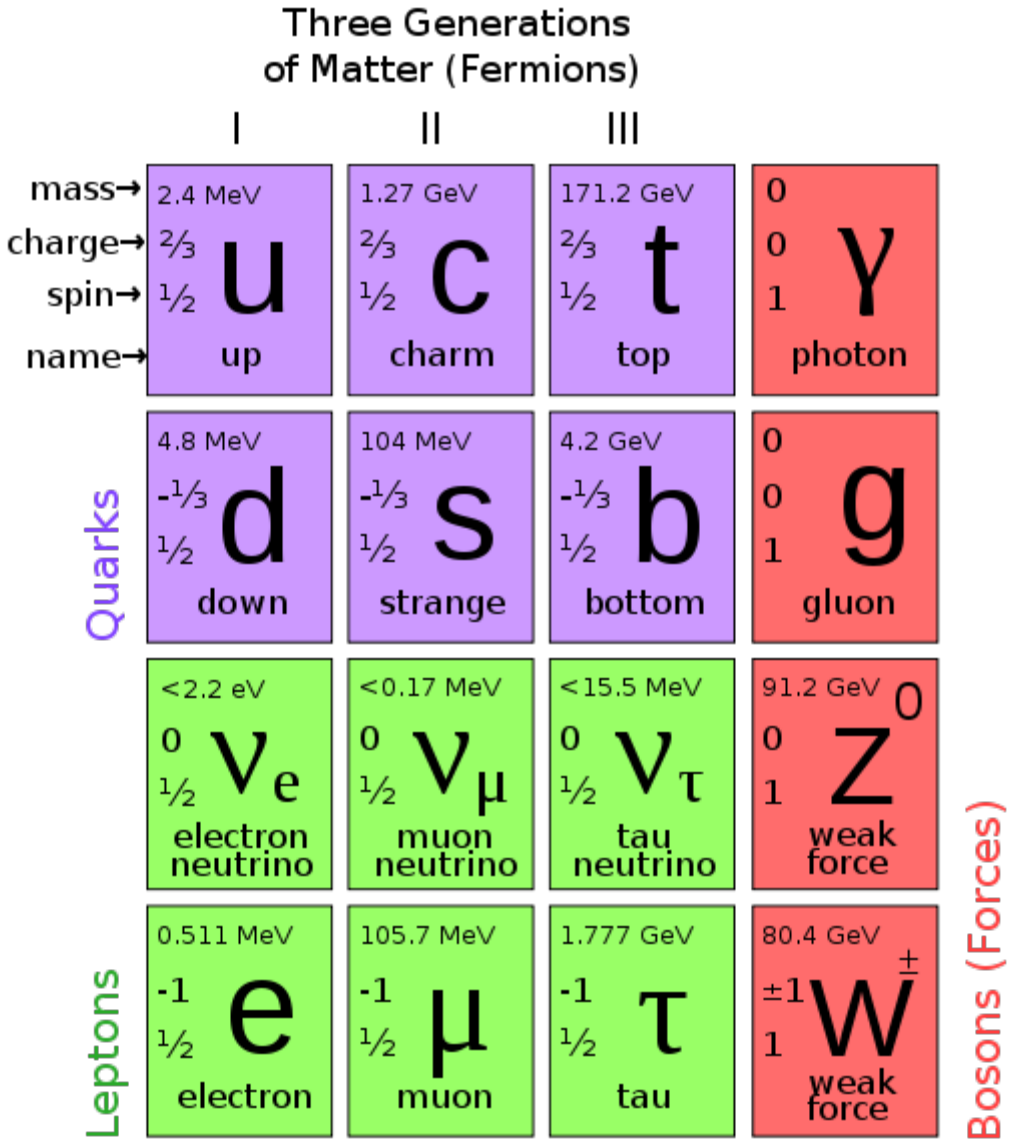
## Particle accelerators

With the development of the particle accelerator during the first half of the twentieth century, physicists began to delve deeper into the properties of subatomic particles. The first successful attempt to accelerate electrons using Electromagnetic induction was made in 1942 by Donald Kerst. His initial betatron reached energies of 2.3 MeV, while subsequent betatrons achieved 300 MeV. In 1947, synchrotron radiation was discovered with a 70 MeV electron synchrotron at General Electric. This radiation was caused by the acceleration of electrons, moving near the speed of light, through a magnetic field.

With a beam energy of 1.5 GeV, the first high-energy particle collider was ADONE, which began operations in 1968. This device accelerated electrons and positrons in opposite directions, effectively doubling the energy of their collision when compared to striking a static target with an electron. The Large Electron-Positron Collider (LEP) at CERN, which was operational from 1989 to 2000, achieved collision energies of 209 GeV and made important measurements for the Standard Model of particle physics.

## *Characteristics*

### Classification




Standard Model of elementary particles. The electron is at lower left.

In the Standard Model of particle physics, electrons belong to the group of subatomic particles called leptons, which are believed to be fundamental or elementary particles. Electrons have the lowest mass of any charged lepton (or electrically charged particle of any type) and belong to the first-generation of fundamental particles. The second and third generation contain charged leptons, the muon and the tau, which are identical to the electron in charge, spin and interactions, but are more massive. Leptons differ from the other basic constituent of matter, the quarks, by their lack of strong interaction. All

members of the lepton group are fermions, because they all have half-odd integer spin; the electron has spin $\frac{1}{2}$.

## Fundamental properties

The invariant mass of an electron is approximately $9.109 \times 10^{-31}$ kilogram, or $5.489 \times 10^{-4}$ atomic mass unit. On the basis of Einstein's principle of mass–energy equivalence, this mass corresponds to a rest energy of 0.511 MeV. The ratio between the mass of a proton and that of an electron is about 1836. Astronomical measurements show that the proton-to-electron mass ratio has held the same value for at least half the age of the universe, as is predicted by the Standard Model.

Electrons have an electric charge of $-1.602 \times 10^{-19}$ coulomb, which is used as a standard unit of charge for subatomic particles. Within the limits of experimental accuracy, the electron charge is identical to the charge of a proton, but with the opposite sign. As the symbol $e$ is used for the elementary charge, the electron is commonly symbolized by $e^-$, where the minus sign indicates the negative charge. The positron is symbolized by $e^+$ because it has the same properties as the electron but with a positive rather than negative charge.

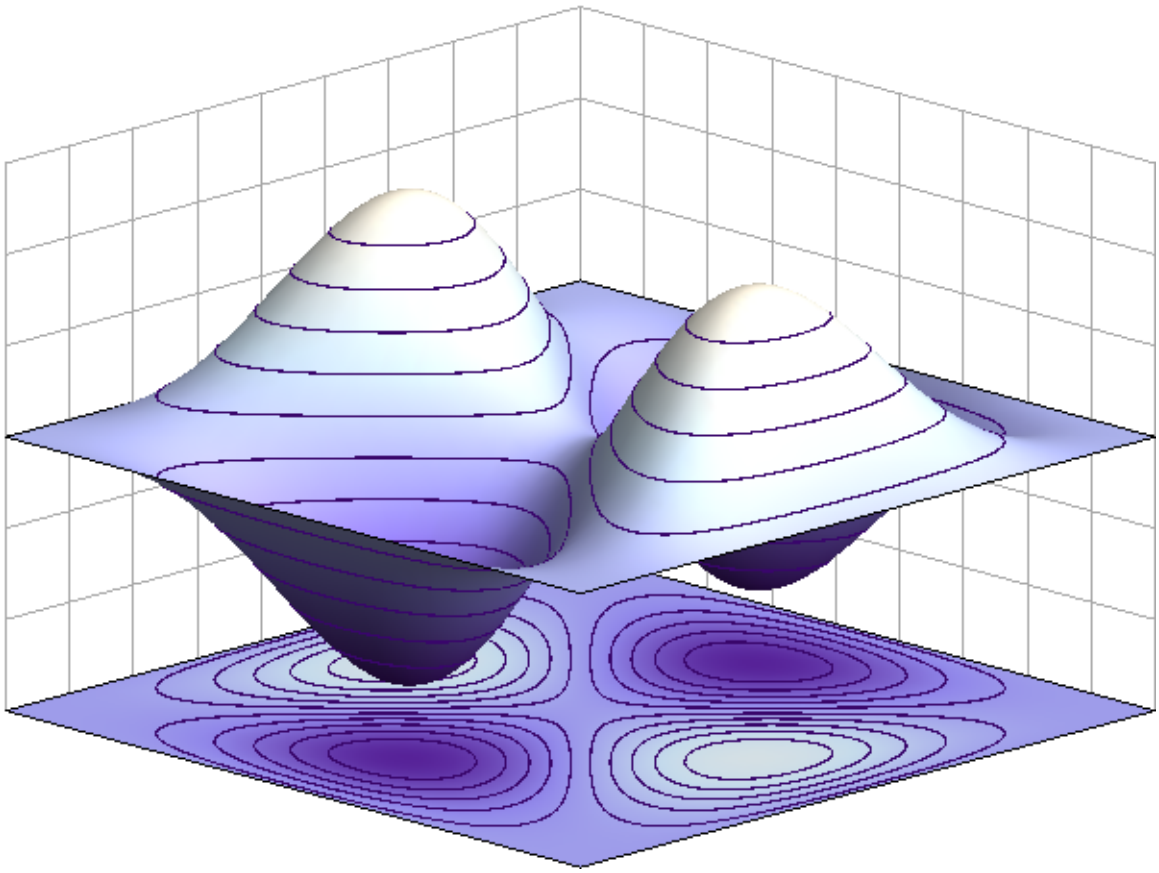
The electron has an intrinsic angular momentum or spin of $\frac{1}{2}$. This property is usually stated by referring to the electron as a spin-$\frac{1}{2}$ particle. For such particles the spin magnitude is $\sqrt{3}\,\frac{1}{2}\,\hbar$. while the result of the measurement of a projection of the spin on any axis can only be $\pm\frac{\hbar}{2}$. In addition to spin, the electron has an intrinsic magnetic moment along its spin axis. It is approximately equal to one Bohr magneton, which is a physical constant equal to $9.27400915(23) \times 10^{-24}$ joules per tesla. The orientation of the spin with respect to the momentum of the electron defines the property of elementary particles known as helicity.

The electron has no known substructure. Hence, it is defined or assumed to be a point particle with a point charge and no spatial extent. Observation of a single electron in a Penning trap shows the upper limit of the particle's radius is $10^{-22}$ meters. There *is* a physical constant called the "classical electron radius", with the much larger value of $2.8179 \times 10^{-15}$ m. However, the terminology comes from a simplistic calculation that ignores the effects of quantum mechanics; in reality, the so-called classical electron radius has little to do with the true fundamental structure of the electron.

There are elementary particles that spontaneously decay into less massive particles. An example is the muon, which decays into an electron, a neutrino and an antineutrino, with a mean lifetime of $2.2 \times 10^{-6}$ seconds. However, the electron is thought to be stable on theoretical grounds: the electron is the least massive particle with non-zero electric charge, so its decay would violate charge conservation. The experimental lower bound for the electron's mean lifetime is $4.6 \times 10^{26}$ years, at a 90% confidence level.

## Quantum properties

As with all particles, electrons can act as waves. This is called the wave–particle duality and can be demonstrated using the double-slit experiment. The wave-like nature of the electron allows it to pass through two parallel slits simultaneously, rather than just one slit as would be the case for a classical particle. In quantum mechanics, the wave-like property of one particle can be described mathematically as a complex-valued function, the wave function, commonly denoted by the Greek letter psi ($\psi$). When the absolute value of this function is squared, it gives the probability that a particle will be observed near a location—a probability density.



Example of an antisymmetric wave function for a quantum state of two identical fermions in a 2-dimensional box. If the particles swap position, the wave function inverts its sign.

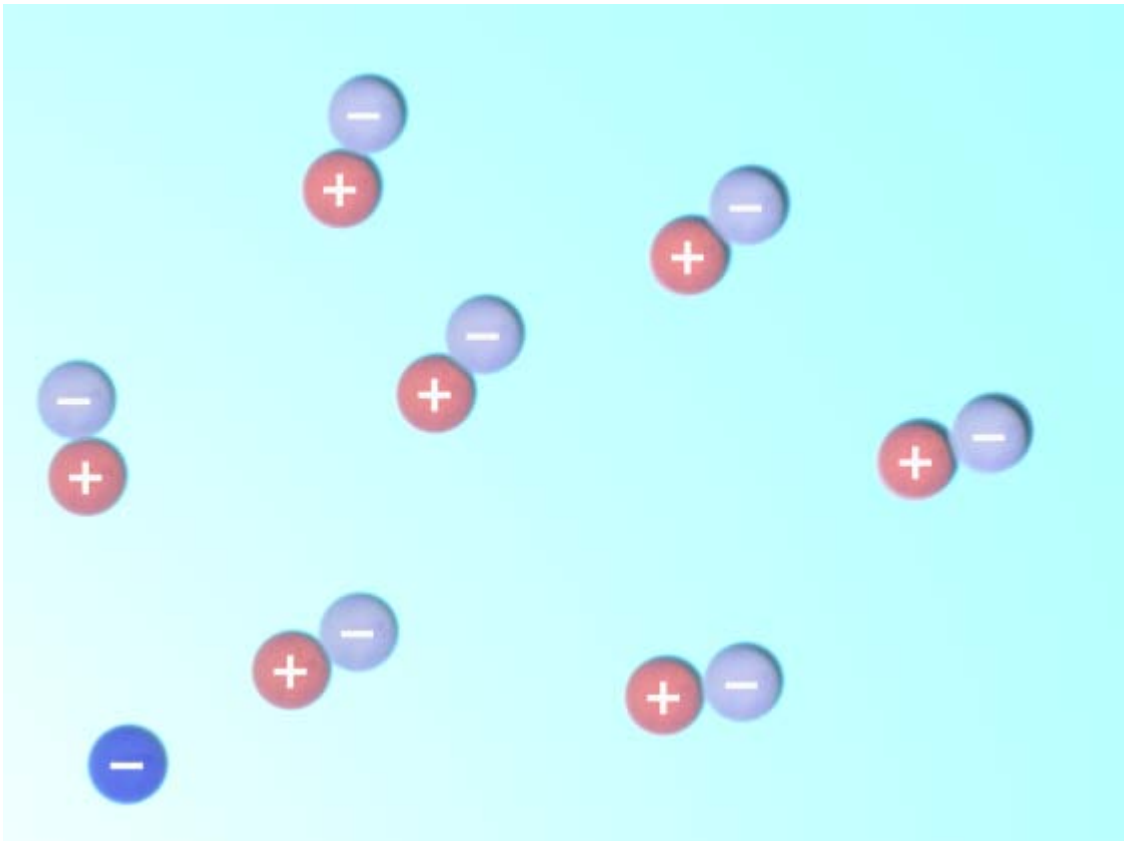
Electrons are identical particles because they cannot be distinguished from each other by their intrinsic physical properties. In quantum mechanics, this means that a pair of interacting electrons must be able to swap positions without an observable change to the state of the system. The wave function of fermions, including electrons, is antisymmetric, meaning that it changes sign when two electrons are swapped; that is, $\psi(r_1, r_2) = -\psi(r_2, r_1)$, where the variables $r_1$ and $r_2$ correspond to the first and second electrons,

respectively. Since the absolute value is not changed by a sign swap, this corresponds to equal probabilities. Bosons, such as the photon, have symmetric wave functions instead.

In the case of antisymmetry, solutions of the wave equation for interacting electrons result in a zero probability that each pair will occupy the same location or state. This is responsible for the Pauli exclusion principle, which precludes any two electrons from occupying the same quantum state. This principle explains many of the properties of electrons. For example, it causes groups of bound electrons to occupy different orbitals in an atom, rather than all overlapping each other in the same orbit.

## Virtual particles

Physicists believe that empty space may be continually creating pairs of virtual particles, such as a positron and electron, which rapidly annihilate each other shortly thereafter. The combination of the energy variation needed to create these particles, and the time during which they exist, fall under the threshold of detectability expressed by the Heisenberg uncertainty relation, $\Delta E \cdot \Delta t \geq \hbar$. In effect, the energy needed to create these virtual particles, $\Delta E$, can be "borrowed" from the vacuum for a period of time, $\Delta t$, so that their product is no more than the reduced Planck constant, $\hbar \approx 6.6\times10^{-16}$ eV·s. Thus, for a virtual electron, $\Delta t$ is at most $1.3\times10^{-21}$ s.



A schematic depiction of virtual electron–positron pairs appearing at random near an electron (at lower left)

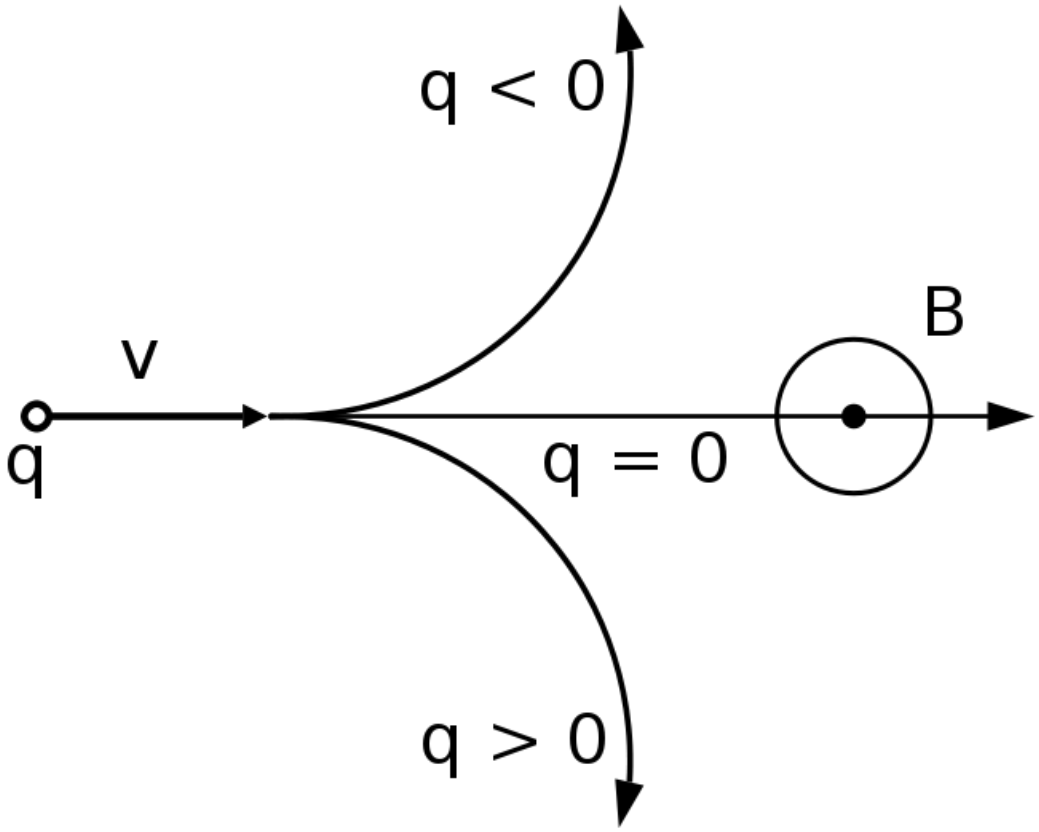
While an electron–positron virtual pair is in existence, the coulomb force from the ambient electric field surrounding an electron causes a created positron to be attracted to the original electron, while a created electron experiences a repulsion. This causes what is called vacuum polarization. In effect, the vacuum behaves like a medium having a dielectric permittivity more than unity. Thus the effective charge of an electron is actually smaller than its true value, and the charge decreases with increasing distance from the electron. This polarization was confirmed experimentally in 1997 using the Japanese TRISTAN particle accelerator. Virtual particles cause a comparable shielding effect for the mass of the electron.

The interaction with virtual particles also explains the small (about 0.1%) deviation of the intrinsic magnetic moment of the electron from the Bohr magneton (the anomalous magnetic moment). The extraordinarily precise agreement of this predicted difference with the experimentally determined value is viewed as one of the great achievements of quantum electrodynamics.

In classical physics, the angular momentum and magnetic moment of an object depend upon its physical dimensions. Hence, the concept of a dimensionless electron possessing these properties might seem inconsistent. The apparent paradox can be explained by the formation of virtual photons in the electric field generated by the electron. These photons cause the electron to shift about in a jittery fashion (known as zitterbewegung), which results in a net circular motion with precession. This motion produces both the spin and the magnetic moment of the electron. In atoms, this creation of virtual photons explains the Lamb shift observed in spectral lines.

## Interaction

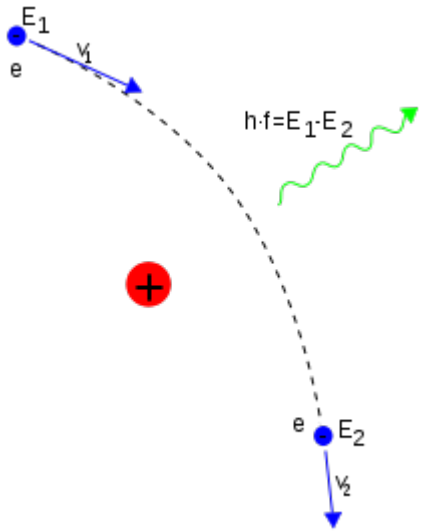
An electron generates an electric field that exerts an attractive force on a particle with a positive charge, such as the proton, and a repulsive force on a particle with a negative charge. The strength of this force is determined by Coulomb's inverse square law. When an electron is in motion, it generates a magnetic field. The Ampère-Maxwell law relates the magnetic field to the mass motion of electrons (the current) with respect to an observer. It is this property of induction which supplies the magnetic field that drives an electric motor. The electromagnetic field of an arbitrary moving charged particle is expressed by the Liénard–Wiechert potentials, which are valid even when the particle's speed is close to that of light (relativistic).




A particle with charge *q* (at left) is moving with velocity *v* through a magnetic field *B* that is oriented toward the viewer. For an electron, *q* is negative so it follows a curved trajectory toward the top.

When an electron is moving through a magnetic field, it is subject to the Lorentz force that exerts an influence in a direction perpendicular to the plane defined by the magnetic field and the electron velocity. This centripetal force causes the electron to follow a helical trajectory through the field at a radius called the gyroradius. The acceleration from this curving motion induces the electron to radiate energy in the form of synchrotron radiation. The energy emission in turn causes a recoil of the electron, known as the Abraham-Lorentz-Dirac force, which creates a friction that slows the electron. This force is caused by a back-reaction of the electron's own field upon itself.

In quantum electrodynamics the electromagnetic interaction between particles is mediated by photons. An isolated electron that is not undergoing acceleration is unable to emit or absorb a real photon; doing so would violate conservation of energy and momentum. Instead, virtual photons can transfer momentum between two charged particles. It is this exchange of virtual photons that, for example, generates the Coulomb force. Energy emission can occur when a moving electron is deflected by a charged particle, such as a proton. The acceleration of the electron results in the emission of Bremsstrahlung radiation.




Here, Bremsstrahlung is produced by an electron *e* deflected by the electric field of an atomic nucleus. The energy change $E_2 - E_1$ determines the frequency *f* of the emitted photon.

An inelastic collision between a photon (light) and a solitary (free) electron is called Compton scattering. This collision results in a transfer of momentum and energy between the particles, which modifies the wavelength of the photon by an amount called the Compton shift. The maximum magnitude of this wavelength shift is $h/m_e c$, which is known as the Compton wavelength. For an electron, it has a value of $2.43 \times 10^{-12}$ m. When the wavelength of the light is long (for instance, the wavelength of the visible light is 0.4–0.7 μm) the wavelength shift becomes negligible. Such interaction between the light and free electrons is called Thomson scattering or Linear Thomson scattering.

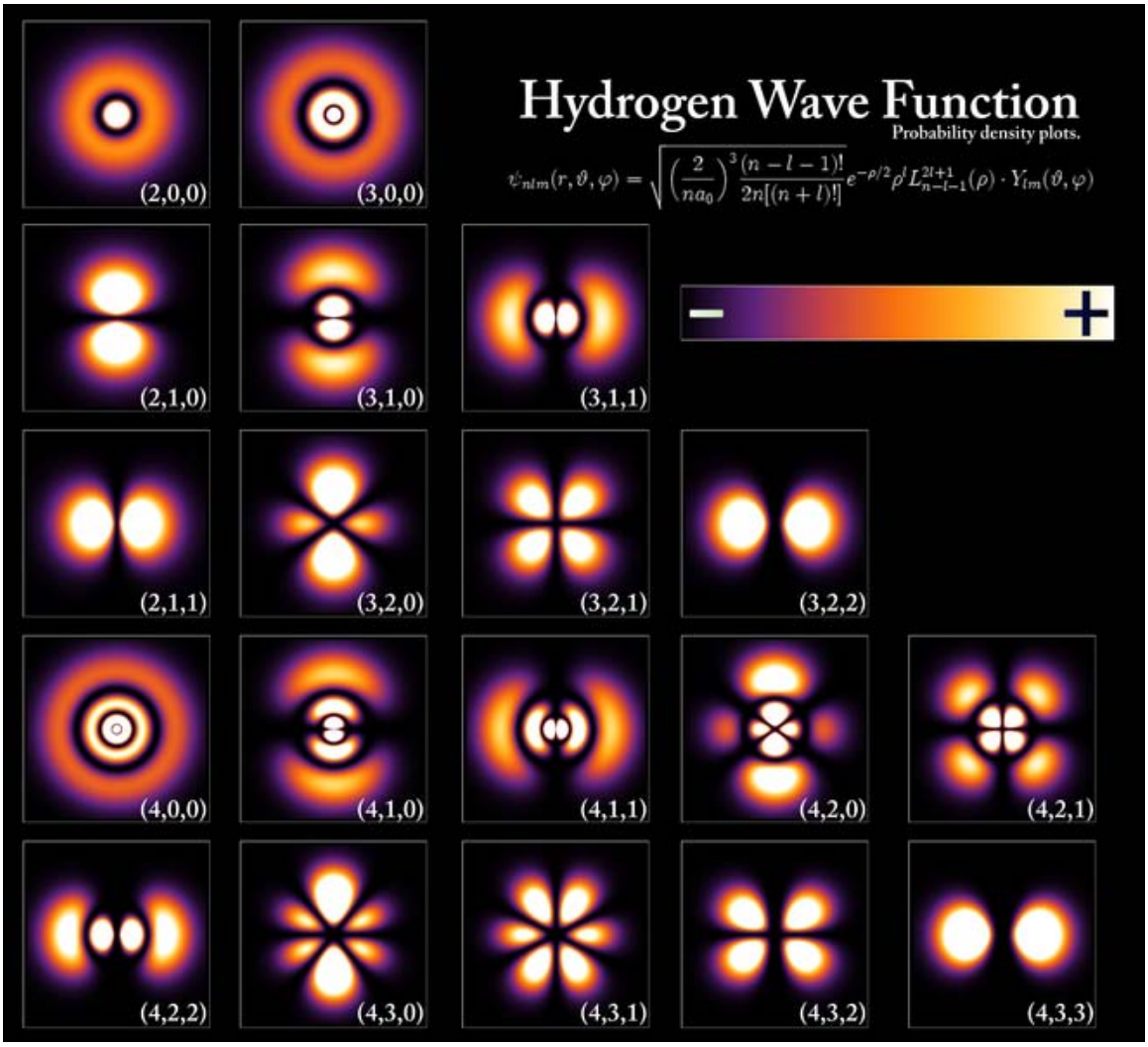
The relative strength of the electromagnetic interaction between two charged particles, such as an electron and a proton, is given by the fine-structure constant. This value is a dimensionless quantity formed by the ratio of two energies: the electrostatic energy of attraction (or repulsion) at a separation of one Compton wavelength, and the rest energy of the charge. It is given by $\alpha \approx 7.297353 \times 10^{-3}$, which is approximately equal to $1/137$.

When electrons and positrons collide, they annihilate each other, giving rise to two or more gamma ray photons. If the electron and positron have negligible momentum, a positronium atom can form before annihilation results in two or three gamma ray photons totalling 1.022 MeV. On the other hand, high-energy photons may transform into an electron and a positron by a process called pair production, but only in the presence of a nearby charged particle, such as a nucleus.

In the theory of electroweak interaction, the left-handed component of electron's wavefunction forms a weak isospin doublet with the electron neutrino. This means that during weak interactions, electron neutrinos behave like electrons. Either member of this doublet can undergo a charged current interaction by emitting or absorbing a W and be

converted into the other member. Charge is conserved during this reaction because the W boson also carries a charge, canceling out any net change during the transmutation. Charged current interactions are responsible for the phenomenon of beta decay in a radioactive atom. Both the electron and electron neutrino can undergo a neutral current interaction via a $Z^0$
exchange, and this is responsible for neutrino-electron elastic scattering.

## Atoms and molecules




Probability densities for the first few hydrogen atom orbitals, seen in cross-section. The energy level of a bound electron determines the orbital it occupies, and the color reflects the probability to find the electron at a given position.

An electron can be *bound* to the nucleus of an atom by the attractive Coulomb force. A system of several electrons bound to a nucleus is called an atom. If the number of electrons is different from the nucleus' electrical charge, such an atom is called an ion. The wave-like behavior of a bound electron is described by a function called an atomic

orbital. Each orbital has its own set of quantum numbers such as energy, angular momentum and projection of angular momentum, and only a discrete set of these orbitals exist around the nucleus. According to the Pauli exclusion principal each orbital can be occupied by up to two electrons, which must differ in their spin quantum number.

Electrons can transfer between different orbitals by the emission or absorption of photons with an energy that matches the difference in potential. Other methods of orbital transfer include collisions with particles, such as electrons, and the Auger effect. In order to escape the atom, the energy of the electron must be increased above its binding energy to the atom. This occurs, for example, with the photoelectric effect, where an incident photon exceeding the atom's ionization energy is absorbed by the electron.

The orbital angular momentum of electrons is quantized. Because the electron is charged, it produces an orbital magnetic moment that is proportional to the angular momentum. The net magnetic moment of an atom is equal to the vector sum of orbital and spin magnetic moments of all electrons and the nucleus. The nuclear magnetic moment is, however, negligible in comparison to the effect from the electrons. The magnetic moments of the electrons that occupy the same orbital (so called, paired electrons) cancel each other out.

The chemical bond between atoms occurs as a result of electromagnetic interactions, as described by the laws of quantum mechanics. The strongest bonds are formed by the sharing or transfer of electrons between atoms, allowing the formation of molecules. Within a molecule, electrons move under the influence of several nuclei, and occupy molecular orbitals; much as they can occupy atomic orbitals in isolated atoms. A fundamental factor in these molecular structures is the existence of electron pairs. These are electrons with opposed spins, allowing them to occupy the same molecular orbital without violating the Pauli exclusion principle (much like in atoms). Different molecular orbitals have different spatial distribution of the electron density. For instance, in bonded pairs (i.e. in the pairs that actually bind atoms together) electrons can be found with the maximal probability in a relatively small volume between the nuclei. On the contrary, in non-bonded pairs electrons are distributed in a large volume around nuclei.

## Conductivity

A lightning discharge consists primarily of a flow of electrons. The electric potential needed for lightning may be generated by a triboelectric effect.

If a body has more or fewer electrons than are required to balance the positive charge of the nuclei, then that object has a net electric charge. When there is an excess of electrons, the object is said to be negatively charged. When there are fewer electrons than the number of protons in nuclei, the object is said to be positively charged. When the number of electrons and the number of protons are equal, their charges cancel each other and the object is said to be electrically neutral. A macroscopic body can develop an electric charge through rubbing, by the triboelectric effect.

Independent electrons moving in vacuum are termed *free* electrons. Electrons in metals also behave as if they were free. In reality the particles that are commonly termed electrons in metals and other solids are quasi-electrons—quasi-particles, which have the same electrical charge, spin and magnetic moment as real electrons but may have a different mass. When free electrons—both in vacuum and metals—move, they produce a net flow of charge called an electric current, which generates a magnetic field. Likewise a current can be created by a changing magnetic field. These interactions are described mathematically by Maxwell's equations.

At a given temperature, each material has an electrical conductivity that determines the value of electric current when an electric potential is applied. Examples of good conductors include metals such as copper and gold, whereas glass and Teflon are poor conductors. In any dielectric material, the electrons remain bound to their respective atoms and the material behaves as an insulator. Most semiconductors have a variable level of conductivity that lies between the extremes of conduction and insulation. On the other hand, metals have an electronic band structure containing partially filled electronic bands. The presence of such bands allows electrons in metals to behave as if they were free or delocalized electrons. These electrons are not associated with specific atoms, so when an electric field is applied, they are free to move like a gas (called Fermi gas) through the material much like free electrons.

Because of collisions between electrons and atoms, the drift velocity of electrons in a conductor is on the order of millimeters per second. However, the speed at which a change of current at one point in the material causes changes in currents in other parts of the material, the velocity of propagation, is typically about 75% of light speed. This occurs because electrical signals propagate as a wave, with the velocity dependent on the dielectric constant of the material.

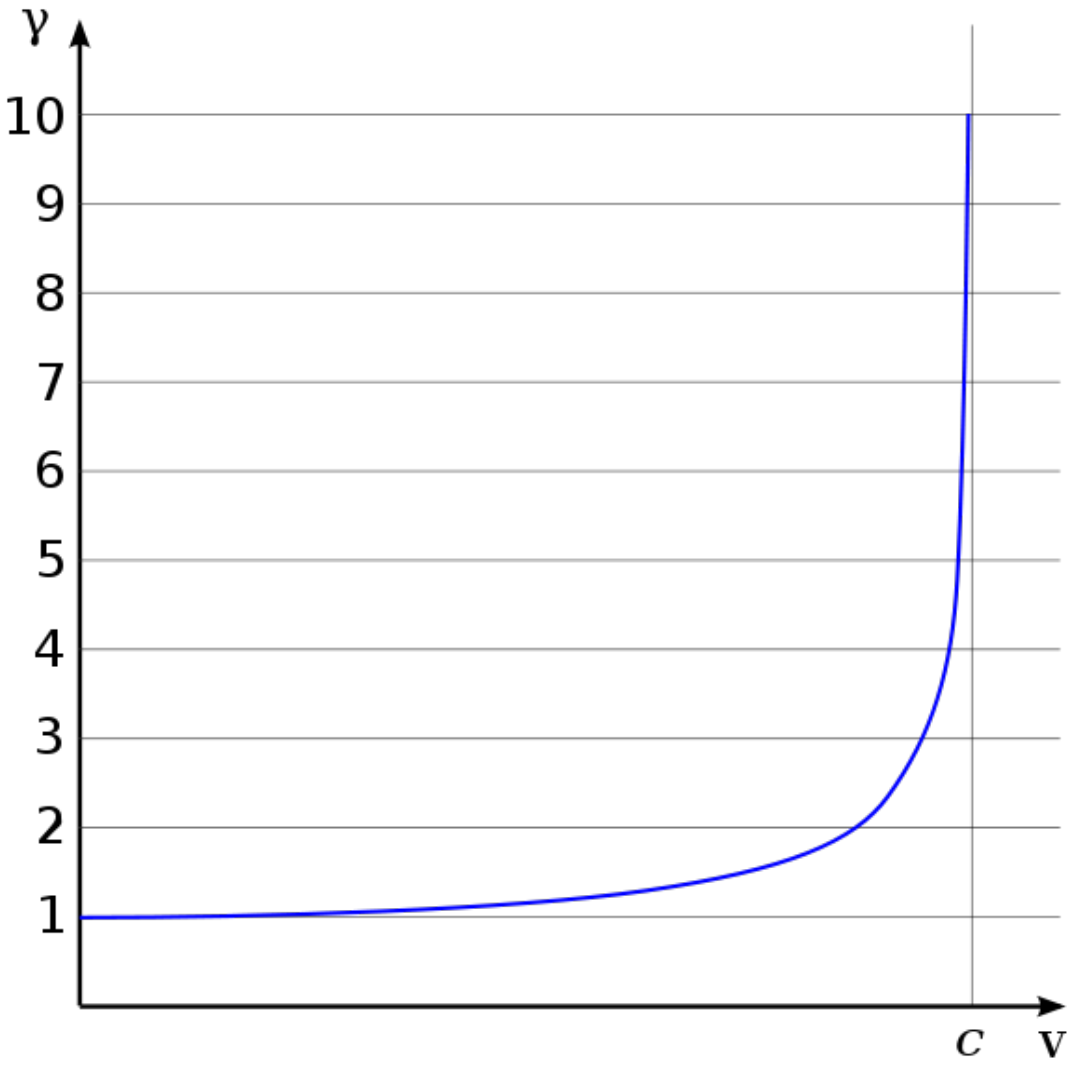
Metals make relatively good conductors of heat, primarily because the delocalized electrons are free to transport thermal energy between atoms. However, unlike electrical conductivity, the thermal conductivity of a metal is nearly independent of temperature. This is expressed mathematically by the Wiedemann-Franz law, which states that the ratio of thermal conductivity to the electrical conductivity is proportional to the temperature. The thermal disorder in the metallic lattice increases the electrical resistivity of the material, producing a temperature dependence for electrical current.

When cooled below a point called the critical temperature, materials can undergo a phase transition in which they lose all resistivity to electrical current, in a process known as superconductivity. In BCS theory, this behavior is modeled by pairs of electrons entering a quantum state known as a Bose–Einstein condensate. These Cooper pairs have their motion coupled to nearby matter via lattice vibrations called phonons, thereby avoiding the collisions with atoms that normally create electrical resistance. (Cooper pairs have a radius of roughly 100 nm, so they can overlap each other.) However, the mechanism by which higher temperature superconductors operate remains uncertain.

Electrons inside conducting solids, which are quasi-particles themselves, when tightly confined at temperatures close to absolute zero, behave as though they had split into two other quasiparticles: spinons and holons. The former carries spin and magnetic moment, while the latter electrical charge.

## Motion and energy

According to Einstein's theory of special relativity, as an electron's speed approaches the speed of light, from an observer's point of view its relativistic mass increases, thereby making it more and more difficult to accelerate it from within the observer's frame of reference. The speed of an electron can approach, but never reach, the speed of light in a vacuum, *c*. However, when relativistic electrons—that is, electrons moving at a speed close to *c*—are injected into a dielectric medium such as water, where the local speed of light is significantly less than *c*, the electrons temporarily travel faster than light in the medium. As they interact with the medium, they generate a faint light called Cherenkov radiation.



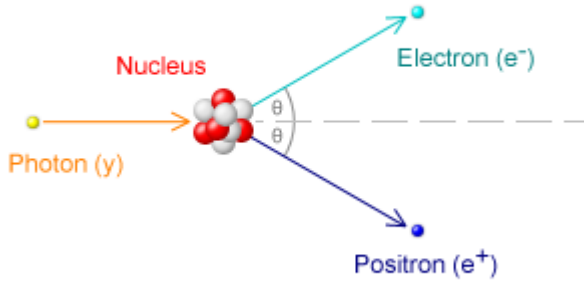

Lorentz factor as a function of velocity. It starts at value 1 and goes to infinity as $v$ approaches $c$.

The effects of special relativity are based on a quantity known as the Lorentz factor, defined as $\gamma = 1/\sqrt{1 - v^2/c^2}$ where $v$ is the speed of the particle. The kinetic energy $K_e$ of an electron moving with velocity $v$ is:

$$K_e = (\gamma - 1)m_e c^2,$$

where $m_e$ is the mass of electron. For example, the Stanford linear accelerator can accelerate an electron to roughly 51 GeV. Since an electron behaves as a wave, at a given velocity it has a characteristic de Broglie wavelength. This is given by $\lambda_e = h/p$ where $h$ is the Planck constant and $p$ is the momentum. For the 51 GeV electron above, the wavelength is about $2.4 \times 10^{-17}$ m, small enough to explore structures well below the size of an atomic nucleus.

## *Formation*




Pair production caused by the collision of a photon with an atomic nucleus

The Big Bang theory is the most widely accepted scientific theory to explain the early stages in the evolution of the Universe. For the first millisecond of the Big Bang, the temperatures were over 10 billion kelvins and photons had mean energies over a million electronvolts. These photons were sufficiently energetic that they could react with each other to form pairs of electrons and positrons,

$$\gamma + \gamma \rightleftharpoons \mathrm{e}^+ + \mathrm{e}^-,$$

where γ is a photon, e+is a positron and e−is an electron. Likewise, positron-electron pairs annihilated each other and emitted energetic photons. An equilibrium between electrons, positrons and photons was maintained during this phase of the evolution of the Universe. After 15 seconds had passed, however, the temperature of the universe dropped below the threshold where electron-positron formation could occur. Most of the surviving electrons and positrons annihilated each other, releasing gamma radiation that briefly reheated the universe.

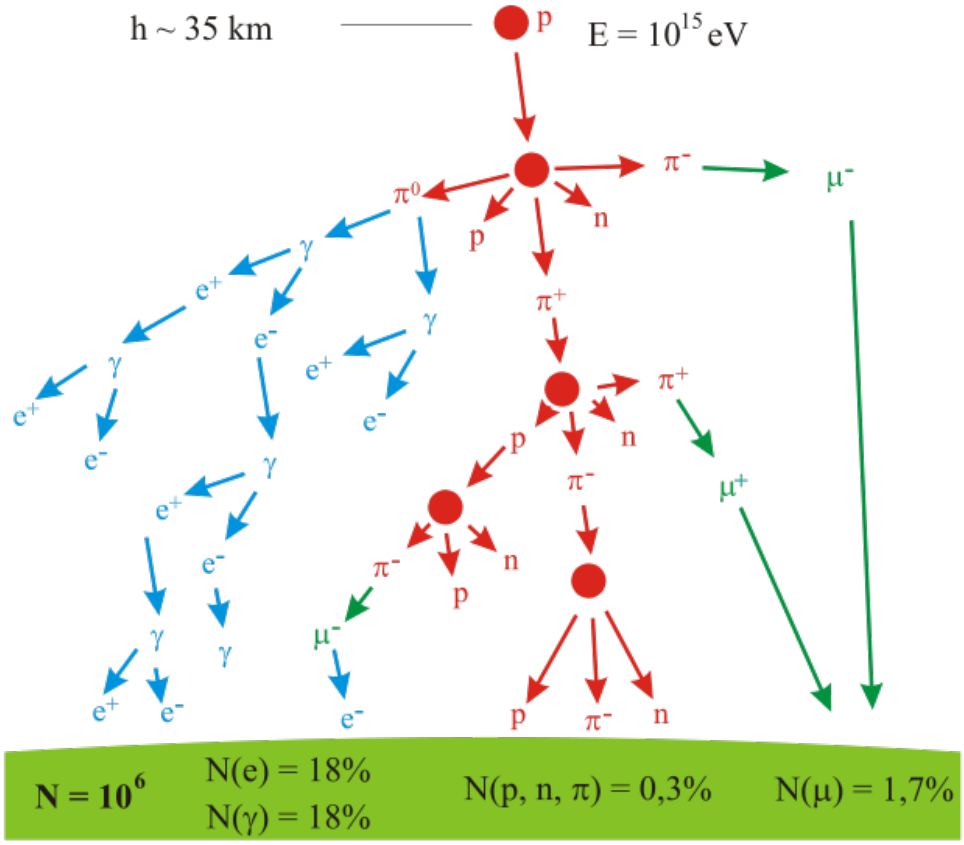
For reasons that remain uncertain, during the process of leptogenesis there was an excess in the number of electrons over positrons. Hence, about one electron in every billion survived the annihilation process. This excess matched the excess of protons over anti-protons, in a condition known as baryon asymmetry, resulting in a net charge of zero for the universe. The surviving protons and neutrons began to participate in reactions with each other—in the process known as nucleosynthesis, forming isotopes of hydrogen and helium, with trace amounts of lithium. This process peaked after about five minutes. Any leftover neutrons underwent negative beta decay with a half-life of about a thousand seconds, releasing a proton and electron in the process,

$$\mathrm{n} \Rightarrow \mathrm{p} + \mathrm{e}^- + \bar{\nu}_\mathrm{e},$$

where n is a neutron, p is a proton and νe is an electron antineutrino. For about the next 300,000–400,000 years, the excess electrons remained too energetic to bind with atomic

nuclei. What followed is a period known as recombination, when neutral atoms were formed and the expanding universe became transparent to radiation.

Roughly one million years after the big bang, the first generation of stars began to form. Within a star, stellar nucleosynthesis results in the production of positrons from the fusion of atomic nuclei. These antimatter particles immediately annihilate with electrons, releasing gamma rays. The net result is a steady reduction in the number of electrons, and a matching increase in the number of neutrons. However, the process of stellar evolution can result in the synthesis of radioactive isotopes. Selected isotopes can subsequently undergo negative beta decay, emitting an electron and antineutrino from the nucleus. An example is the cobalt-60 ($^{60}$Co) isotope, which decays to form nickel-60 ($^{60}$Ni).




An extended air shower generated by an energetic cosmic ray striking the Earth's atmosphere

At the end of its lifetime, a star with more than about 20 solar masses can undergo gravitational collapse to form a black hole. According to classical physics, these massive stellar objects exert a gravitational attraction that is strong enough to prevent anything, even electromagnetic radiation, from escaping past the Schwarzschild radius. However, it is believed that quantum mechanical effects may allow Hawking radiation to be emitted

at this distance. Electrons (and positrons) are thought to be created at the event horizon of these stellar remnants.

When pairs of virtual particles (such as an electron and positron) are created in the vicinity of the event horizon, the random spatial distribution of these particles may permit one of them to appear on the exterior; this process is called quantum tunneling. The gravitational potential of the black hole can then supply the energy that transforms this virtual particle into a real particle, allowing it to radiate away into space. In exchange, the other member of the pair is given negative energy, which results in a net loss of mass-energy by the black hole. The rate of Hawking radiation increases with decreasing mass, eventually causing the black hole to evaporate away until, finally, it explodes.

Cosmic rays are particles traveling through space with high energies. Energy events as high as $3.0\times10^{20}$ eV have been recorded. When these particles collide with nucleons in the Earth's atmosphere, a shower of particles is generated, including pions. More than half of the cosmic radiation observed from the Earth's surface consists of muons. The particle called a muon is a lepton which is produced in the upper atmosphere by the decay of a pion. A muon, in turn, can decay to form an electron or positron. Thus, for the negatively charged pion $\pi^-$,

$$\pi^- \rightarrow \mu^- + \bar{\nu_\mu},$$
$$\mu^- \rightarrow \mathrm{e}^- + \bar{\nu}_\mathrm{e} + \nu_\mu,$$

where $\mu^-$ is a muon and $\nu_\mu$ is a muon neutrino.

## *Observation*

Aurorae are mostly caused by energetic electrons precipitating into the atmosphere.

Remote observation of electrons requires detection of their radiated energy. For example, in high-energy environments such as the corona of a star, free electrons form a plasma that radiates energy due to Bremsstrahlung. Electron gas can undergo plasma oscillation, which is waves caused by synchronized variations in electron density, and these produce energy emissions that can be detected by using radio telescopes.

The frequency of a photon is proportional to its energy. As a bound electron transitions between different energy levels of an atom, it will absorb or emit photons at characteristic frequencies. For instance, when atoms are irradiated by a source with a broad spectrum, distinct absorption lines will appear in the spectrum of transmitted radiation. Each element or molecule displays a characteristic set of spectral lines, such as the hydrogen spectral series. Spectroscopic measurements of the strength and width of these lines allow the composition and physical properties of a substance to be determined.

In laboratory conditions, the interactions of individual electrons can be observed by means of particle detectors, which allow measurement of specific properties such as energy, spin and charge. The development of the Paul trap and Penning trap allows charged particles to be contained within a small region for long durations. This enables precise measurements of the particle properties. For example, in one instance a Penning trap was used to contain a single electron for a period of 10 months. The magnetic

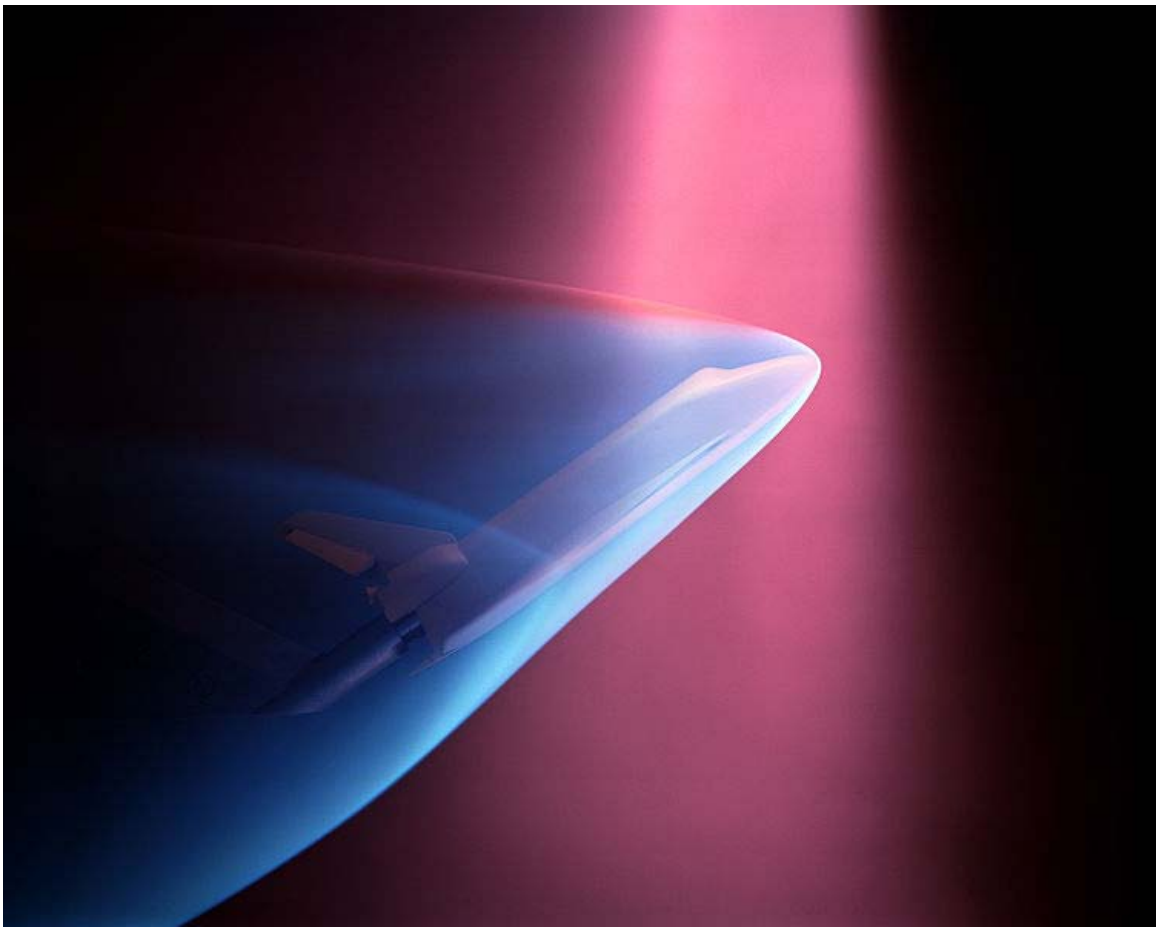
moment of the electron was measured to a precision of eleven digits, which, in 1980, was a greater accuracy than for any other physical constant.

The first video images of an electron's energy distribution were captured by a team at Lund University in Sweden, February 2008. The scientists used extremely short flashes of light, called attosecond pulses, which allowed an electron's motion to be observed for the first time.

The distribution of the electrons in solid materials can be visualized by angle resolved photoemission spectroscopy (ARPES). This technique employs the photoelectric effect to measure the reciprocal space—a mathematical representation of periodic structures that is used to infer the original structure. ARPES can be used to determine the direction, speed and scattering of electrons within the material.

## *Plasma applications*

### Particle beams



During a NASA wind tunnel test, a model of the Space Shuttle is targeted by a beam of electrons, simulating the effect of ionizing gases during re-entry.

Electron beams are used in welding, which allows energy densities up to 10$^{7}$ W·cm$^{-2}$ across a narrow focus diameter of 0.1–1.3 mm and usually does not require a filler material. This welding technique must be performed in a vacuum, so that the electron beam does not interact with the gas prior to reaching the target, and it can be used to join conductive materials that would otherwise be considered unsuitable for welding.

Electron beam lithography (EBL) is a method of etching semiconductors at resolutions smaller than a micron. This technique is limited by high costs, slow performance, the need to operate the beam in the vacuum and the tendency of the electrons to scatter in solids. The last problem limits the resolution to about 10 nm. For this reason, EBL is primarily used for the production of small numbers of specialized integrated circuits.

Electron beam processing is used to irradiate materials in order to change their physical properties or sterilize medical and food products. In radiation therapy, electron beams are generated by linear accelerators for treatment of superficial tumors. Because an electron beam only penetrates to a limited depth before being absorbed, typically up to 5 cm for electron energies in the range 5–20 MeV, electron therapy is useful for treating skin lesions such as basal cell carcinomas. An electron beam can be used to supplement the treatment of areas that have been irradiated by X-rays.

Particle accelerators use electric fields to propel electrons and their antiparticles to high energies. As these particles pass through magnetic fields, they emit synchrotron radiation. The intensity of this radiation is spin dependent, which causes polarization of the electron beam—a process known as the Sokolov–Ternov effect. The polarized electron beams can be useful for various experiments. Synchrotron radiation can also be used for cooling the electron beams, which reduces the momentum spread of the particles. Once the particles have accelerated to the required energies, separate electron and positron beams are brought into collision. The resulting energy emissions are observed with particle detectors and are studied in particle physics.

## Imaging

Low-energy electron diffraction (LEED) is a method of bombarding a crystalline material with a collimated beam of electrons, then observing the resulting diffraction patterns to determine the structure of the material. The required energy of the electrons is typically in the range 20–200 eV. The reflection high energy electron diffraction (RHEED) technique uses the reflection of a beam of electrons fired at various low angles to characterize the surface of crystalline materials. The beam energy is typically in the range 8–20 keV and the angle of incidence is 1–4°.

The electron microscope directs a focused beam of electrons at a specimen. As the beam interacts with the material, some electrons change their properties, such as movement direction, angle, relative phase and energy. By recording these changes in the electron beam, microscopists can produce atomically resolved image of the material. In blue light, conventional optical microscopes have a diffraction-limited resolution of about 200 nm. By comparison, electron microscopes are limited by the de Broglie wavelength of the

electron. This wavelength, for example, is equal to 0.0037 nm for electrons accelerated across a 100,000-volt potential. The Transmission Electron Aberration-corrected Microscope is capable of sub-0.05 nm resolution, which is more than enough to resolve individual atoms. This capability makes the electron microscope a useful laboratory instrument for high resolution imaging. However, electron microscopes are expensive instruments that are costly to maintain.

There are two main types of electron microscopes: transmission and scanning. Transmission electron microscopes function in a manner similar to overhead projector, with a beam of electrons passing through a slice of material then being projected by lenses on a photographic slide or a charge-coupled device. In scanning electron microscopes, the image is produced by rastering a finely focused electron beam, as in a TV set, across the studied sample. The magnifications range from 100× to 1,000,000× or higher for both microscope types. The scanning tunneling microscope uses quantum tunneling of electrons from a sharp metal tip into the studied material and can produce atomically resolved images of its surface.

## Other

In the free electron laser (FEL), a relativistic electron beam is passed through a pair of undulators containing arrays of dipole magnets, whose fields are oriented in alternating directions. The electrons emit synchrotron radiation, which, in turn, coherently interacts with the same electrons. This leads to the strong amplification of the radiation field at the resonance frequency. FEL can emit a coherent high-brilliance electromagnetic radiation with a wide range of frequencies, from microwaves to soft X-rays. These devices can be used in the future for manufacturing, communication and various medical applications, such as soft tissue surgery.

Electrons are at the heart of cathode ray tubes, which are used extensively as display devices in laboratory instruments, computer monitors and television sets. In a photomultiplier tube, every photon striking the photocathode initiates an avalanche of electrons that produces a detectable current pulse. Vacuum tubes use the flow of electrons to manipulate electrical signals, and they played a critical role in the development of electronics technology. However, they have been largely supplanted by solid-state devices such as the transistor.

Chapter-4

# High-k Dielectric and Low-k Dielectric

## High-k dielectric

The term **high-κ dielectric** refers to a material with a high dielectric constant κ (as compared to silicon dioxide) used in semiconductor manufacturing processes which replaces the silicon dioxide gate dielectric. The implementation of high-κ gate dielectrics is one of several strategies developed to allow further miniaturization of microelectronic components, colloquially referred to as extending Moore's Law.

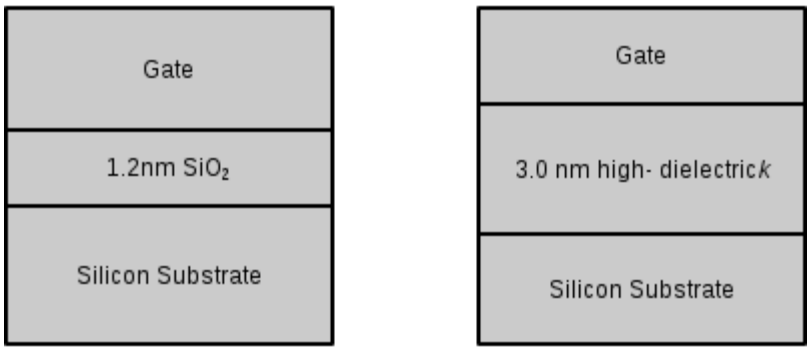
### *Need for high-κ materials*

Silicon dioxide has been used as a gate oxide material for decades. As transistors have decreased in size, the thickness of the silicon dioxide gate dielectric has steadily decreased to increase the gate capacitance and thereby drive current and device performance. As the thickness scales below 2 nm, leakage currents due to tunneling increase drastically, leading to unwieldy power consumption and reduced device reliability. Replacing the silicon dioxide gate dielectric with a high-κ material allows increased gate capacitance without the concomitant leakage effects.

### First principles

The gate oxide in a MOSFET can be modeled as a parallel plate capacitor. Ignoring quantum mechanical and depletion effects from the Si substrate and gate, the capacitance *C* of this parallel plate capacitor is given by

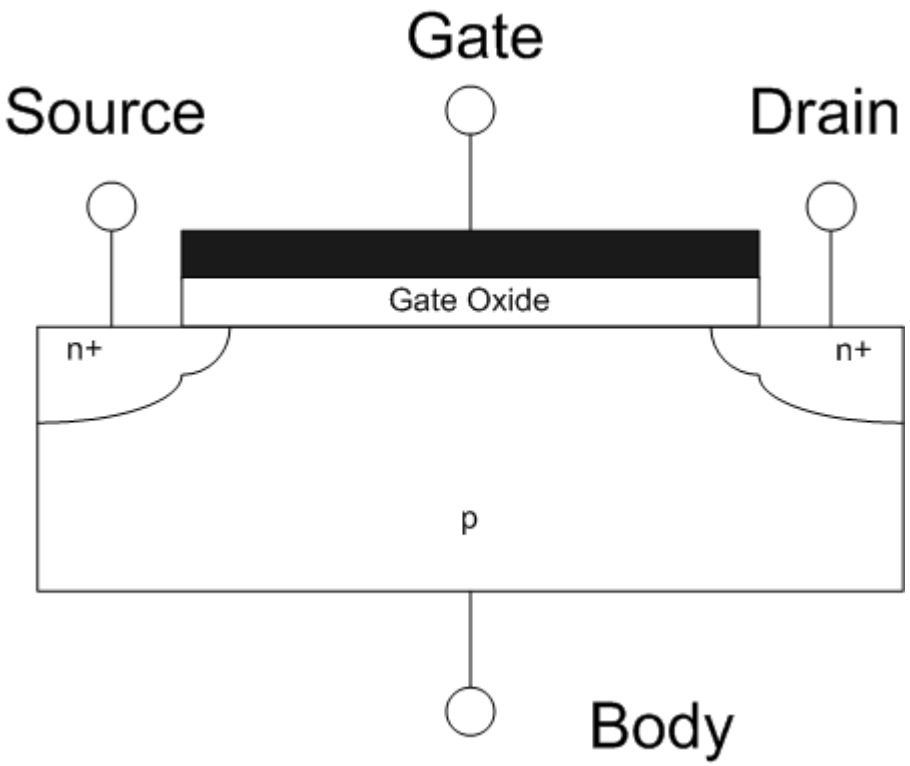
$$C = \frac{\kappa \varepsilon_0 A}{t}$$




Existing 90nm Process
Capacitance = 1x
Leakage Current = 1x

A potential high- process *k*
Capacitance = 1.6x
Leakage Current = 0.01x

Conventional silicon dioxide gate dielectric structure compared to a potential high-k dielectric structure




Cross-section of an NMOS transistor showing the gate oxide dielectric

Where

- $A$ is the capacitor area
- $\kappa$ is the relative dielectric constant of the material (3.9 for silicon dioxide)
- $\varepsilon_0$ is the permittivity of free space
- $t$ is the thickness of the capacitor oxide insulator

Since leakage limitation constrains further reduction of $t$, an alternative method to increase gate capacitance is alter κ by replacing silicon dioxide with a high-κ material. In such a scenario, a thicker gate layer might be used which can reduce the leakage current flowing through the structure as well as improving the gate dielectric reliability.

### Gate capacitance impact on drive current

The drive current $I_D$ for a MOSFET can be written (using the gradual channel approximation) as

$$I_{D,Sat} = \frac{W}{L}\mu\, C_{inv}\frac{(V_G - V_{th})^2}{2}$$

Where

- $W$ is the width of the transistor channel
- $L$ is the channel length
- μ is the channel carrier mobility (assumed constant here)
- $C_{inv}$ is the capacitance density associated with the gate dielectric when the underlying channel is in the inverted state
- $V_G$ is the voltage applied to the transistor gate
- $V_D$ is the voltage applied to the transistor drain
- $V_{th}$ is the threshold voltage

The term $V_G - V_{th}$ is limited in range due to reliability and room temperature operation constraints, since a too large $V_G$ would create an undesirable, high electric field across the oxide. Furthermore, $V_{th}$ cannot easily be reduced below about 200 mV, because leakage currents due to increased oxide leakage (that is, assuming high-κ dielectrics are not available) and subthreshold conduction raise stand-by power consumption to unacceptable levels. Thus, according to this simplified list of factors, an increased $I_{D,sat}$ requires a reduction in the channel length or an increase in the gate dielectric capacitance.

## *Materials and considerations*

Replacing the silicon dioxide gate dielectric with another material adds complexity to the manufacturing process. Silicon dioxide can be formed by oxidizing the underlying silicon, ensuring a uniform, conformal oxide and high interface quality. As a consequence, development efforts have focused on finding a material with a requisitely high dielectric constant that can be easily integrated into a manufacturing process. Other key considerations include band alignment to silicon (which may alter leakage current), film morphology, thermal stability, maintenance of a high mobility of charge carriers in the channel and minimization of electrical defects in the film/interface. Materials which have received considerable attention are hafnium silicate, zirconium silicate, hafnium dioxide and zirconium dioxide, typically deposited using atomic layer deposition.

It is expected that defect states in the high-k dielectric can influence its electrical properties. Defect states can be measured for example by using zero-bias thermally stimulated current, zero-temperature-gradient zero-bias thermally stimulated current spectroscopy, or inelastic electron tunneling spectroscopy (IETS).

### *Use in industry*

The industry has employed oxynitride gate dielectrics since the 1990s, wherein a conventionally formed silicon oxide dielectric is infused with a small amount of nitrogen. The nitride content subtly raises the dielectric constant and is thought to offer other advantages, such as resistance against dopant diffusion through the gate dielectric.

In early 2007, Intel announced the deployment of hafnium-based high-k dielectrics in conjunction with a metallic gate for components built on 45 nanometer technologies, and has shipped it in the 2007 processor series codenamed Penryn. At the same time, IBM announced plans to transition to high-k materials, also hafnium-based, for some products in 2008. While not identified, it is most likely the dielectrics used by these companies are some form of nitrided hafnium silicates (HfSiON). $HfO_2$ and HfSiO are susceptible to crystallization during dopant activation annealing. NEC Electronics has also announced the use of a HfSiON dielectric in their 55 nm *UltimateLowPower* technology. However, even HfSiON is susceptible to trap-related leakage currents, which tend to increase with stress over device lifetime. The higher the hafnium concentration, the more severe the issue. However, there is no absolute guarantee that hafnium will be the basis of future high-k dielectrics. The 2006 ITRS roadmap predicts the implementation of high-k materials to be commonplace in the industry by 2010.

# Low-k dielectric

In semiconductor manufacturing, a **low-κ dielectric** is a material with a small dielectric constant relative to silicon dioxide. Although the proper symbol for the dielectric constant is the Greek letter κ (kappa), in conversation such materials are referred to as being "low-k" (low-kay) rather than "low-κ" (low-kappa). Low-κ dielectric material implementation is one of several strategies used to allow continued scaling of microelectronic devices, colloquially referred to as extending Moore's law. In digital circuits, insulating dielectrics separate the conducting parts (wire interconnects and transistors) from one another. As components have scaled and transistors have got closer together, the insulating dielectrics have thinned to the point where charge build up and crosstalk adversely affect the performance of the device. Replacing the silicon dioxide with a low-κ dielectric of the same thickness reduces parasitic capacitance, enabling faster switching speeds and lower heat dissipation.

### *Low-κ materials*

The dielectric constant of $SiO_2$, the insulating material used in silicon chips, is 3.9. This number is the ratio of the permittivity of $SiO_2$ divided by permittivity of vacuum,

$\varepsilon_{SiO2}/\varepsilon_0$,where $\varepsilon_0 = 8.854\times10^{-6}$ pF/μm . There are many materials with lower dielectric constants but few of them can be suitably integrated into a manufacturing process. Development efforts have focused primarily on three classes of materials:

## Fluorine-doped silicon dioxide

By doping $SiO_2$ with fluorine to produce fluorinated silica glass, the dielectric constant is lowered from 3.9 to 3.5.

## Carbon-doped silicon dioxide

By doping $SiO_2$ with carbon, one can lower the dielectric constant to 3.0. Major products of carbon doped silicon dioxide include **Black Diamond** from Applied Materials, **Aurora** from ASM International N.V. The Aurora is the low-κ material used in Intel 90 nm, 65 nm and 45 nm lines, while the Black Diamond controlled about 80% of low-κ material market. Novellus Systems' **Coral** also falls in this category.

## Porous silicon dioxide

Various methods may be employed to create large voids or pores in a silicon dioxide dielectric. Air has a dielectric constant of roughly 1.0005, thus the dielectric constant of the porous material may be reduced by increasing the porosity of the film. Dielectric constants lower than 2.0 have been reported. Integration difficulties related to porous silicon dioxide implementation include low mechanical strength and difficult integration with etch and polish processes.

## Porous carbon-doped silicon dioxide

By UV curing, floating methyl group in carbon doped silicon dioxide can be eliminated and pores can be introduced to the carbon doped silicon dioxide low-κ materials. The technology for producing porous dielectric layers with very low dielectric constant was developed by A. Grill, V.V. Patel, S.M. Gates in the late 1990's at IBM (US6312793-International Business Machines]. While all materials used for the deposition of low-k dielectric layers are in the public domain and therefore not patentable, the technique of making such layers was rapidly adopted by the semiconductor manufacturing equipment industry and commercialised as an essential application supplied with the equipment. Examples include Applied Materials' **Black Diamond II** technology and ASM International's **Aurora 2.7** and **Aurora ULK** process.

## Spin-on organic polymeric dielectrics

Polymeric dielectrics are generally deposited by a spin-on approach, such as those traditionally used to deposit photoresist, rather than chemical vapor deposition. Integration difficulties include low mechanical strength and thermal stability. **SiLK** from Dow Chemical is a well known example of low-κ material in this category. Other spin-on

organic low-κ include polyimide, polynorbornenes, Benzocyclobutene, PTFE, SU-8 2000.

## Porous SiLK

By introducing pores into the SiLK resin, the dielectric constant value can be lowered to 2.2.

## Spin-on silicone based polymeric dielectric

There are two kinds of silicone based polymeric dielectric materials, hydrogen silsesquioxane (HSQ) and methylsilsesquioxane (MSQ).

Chapter-5

# Semiconductor Device Defects

# Catastrophic optical damage

**Catastrophic optical damage** (**COD**), or **catastrophic optical mirror damage** (**COMD**), is a failure mode of high-power semiconductor lasers. It occurs when the semiconductor junction is overloaded by exceeding its power density and absorbs too much of the produced light energy, leading to melting and recrystallization of the semiconductor material at the facets of the laser. This is often colloquially referred to as "blowing the diode." The affected area contains a large number of lattice defects, negatively affecting its performance. If the affected area is sufficiently large, it can be observable under optical microscope as darkening of the laser facet, and/or as presence of cracks and grooves. The damage can occur within a single laser pulse, in less than a millisecond. The time to COD is inversely proportional to the power density.

Catastrophic optical damage is one of the limiting factors in increasing performance of semiconductor lasers. It is the primary failure mode for AlGaInP/AlGaAs red lasers.

Short-wavelength lasers are more susceptible to COD than long-wavelength ones.

The typical values for COD in industrial products range between 12 and 20 MW/cm$^2$.

## *Causes and mechanisms*

At the edge of a diode laser, where light is emitted, a mirror is traditionally formed by cleaving the semiconductor wafer to form a specularly reflecting plane. This approach is facilitated by the weakness of the [110] crystallographic plane in III-V semiconductor crystals (such as GaAs, InP, GaSb, etc.) compared to other planes. A scratch made at the edge of the wafer and a slight bending force causes a nearly atomically perfect mirror-like cleavage plane to form and propagate in a straight line across the wafer.

But it so happens that the atomic states at the cleavage plane are altered (compared to their bulk properties within the crystal) by the termination of the perfectly periodic lattice at that plane. Surface states at the cleaved plane have energy levels within the (otherwise forbidden) band gap of the semiconductor.

The absorbed light causes generation of electron-hole pairs. These can lead to breaking of chemical bonds on the crystal surface followed by oxidation, or to release of heat by nonradiative recombination. The oxidized surface then shows increased absorption of the laser light, which further accelerates its degradation. The oxidation is especially problematic for semiconductor layers containing aluminium.

Essentially, as a result when light propagates through the cleavage plane and transits to free space from within the semiconductor crystal, a fraction of the light energy is absorbed by the surface states whence it is converted to heat by phonon-electron interactions. This heats the cleaved mirror. In addition the mirror may heat simply because the edge of the diode laser—which is electrically pumped—is in less-than-perfect contact with the mount that provides a path for heat removal. The heating of the mirror causes the band gap of the semiconductor to shrink in the warmer areas. The band gap shrinkage brings more electronic band-to-band transitions into alignment with the photon energy causing yet more absorption. This is thermal runaway, a form of positive feedback, and the result can be melting of the facet, known as *catastrophic optical damage*, or COD.

Deterioration of the laser facets with aging and effects of the environment (erosion by water, oxygen, etc.) increases light absorption by the surface, and decreases the COD threshold. A sudden catastrophic failure of the laser due to COD then can occur after many thousands hours in service.

## *Improvements*

One of the methods of increasing the COD threshold in AlGaInP laser structures is the sulfur treatment, which replaces the oxides at the laser facet with chalcogenide glasses. This decreases the recombination velocity of the surface states.

Reduction of recombination velocity of surface states can be also achieved by cleaving the crystals in ultrahigh vacuum and immediate deposition of a suitable passivation layer.

A thin layer of aluminium can be deposited over the surface, for gettering the oxygen.

Another approach is doping of the surface, increasing the band gap and decreasing absorption of the lasing wavelength, shifting the absorption maximum several nanometers up.

Current crowding near the mirror area can be avoided by prevention of injecting charge carriers near the mirror region. This is achieved by depositing the electrodes away from the mirror, at least several carrier diffusion distances.

Energy density on the surface can be reduced by employing a waveguide broadening the optical cavity, so the same amount of energy exits through a larger area. Energy density of 15–20 $MW/cm^2$ corresponding to 100 mW per micrometer of stripe width are now

achievable. A wider laser stripe can be used for higher output power, for the cost of transverse mode oscillations and therefore worsening of spectral and spatial beam quality.

In the 1970s, this problem, which is particularly nettlesome for GaAs-based lasers emitting between 1 μm and 0.630 μm wavelengths (less so for InP based lasers used for long-haul telecommunications which emit between 1.3 μm and 2 μm), was identified. Michael Ettenberg, a researcher and later Vice President at RCA Laboratories' David Sarnoff Research Center in Princeton, New Jersey, devised a solution. A thin layer of aluminum oxide was deposited on the facet. If the aluminum oxide thickness is chosen correctly, it functions as an anti-reflective coating, reducing reflection at the surface. This alleviated the heating and COD at the facet.

Since then, various other refinements have been employed. One approach is to create a so-called non-absorbing mirror (NAM) such that the final 10 μm or so before the light emits from the cleaved facet are rendered non-absorbing at the wavelength of interest. Such lasers are called **window lasers**.

In the very early 1990s, SDL, Inc. began supplying high power diode lasers with good reliability characteristics. CEO Donald Scifres and CTO David Welch presented new reliability performance data at, e.g., SPIE Photonics West conferences of the era. The methods used by SDL to defeat COD were considered to be highly proprietary and have still not been disclosed publicly as of June, 2006.

In the mid-1990s IBM Research (Ruschlikon, Switzerland) announced that it had devised its so-called "E2 process" which conferred extraordinary resistance to COD in GaAs-based lasers. This process, too, has never been disclosed as of June, 2006.

# Hot carriers injection

**Hot carriers injection** (**HCI**) is the phenomenon in solid-state or semiconductor electronic devices where either an electron or a “hole” gains sufficient kinetic energy to overcome a potential barrier necessary to break an interface state. The term hot electron comes from the effective temperature used to model carrier density (with fermi dirac function). One should not think that the term "hot" refers to the actual temperature of the MOS transistor. At higher temperatures, the mean free path (distance between two collisions with atoms in the substrate) is shorter, which decreases the energy gain by the carrier. As a result, Hot Carrier Degradation is more important at low temperature.

## *HCI and CMOS Semiconductor Technology*

### Semiconductor Physics of HCI

The term “hot carrier injection” usually refers to the effect in MOSFETs, where a carrier is injected from the conducting channel in the silicon substrate to the gate dielectric, which usually is made of silicon dioxide ($SiO_2$).

To become "hot" and enter the conduction band of $SiO_2$, an electron must gain a kinetic energy of 3.3 eV. For holes, the valence band offset in this case dictates they must have a kinetic energy of 4.6 eV.

When electrons are accelerated in the channel, they gain energy along the mean free path. This energy is lost in two different ways:

1. The carrier hit an atom in the substrate. Then the collision create a cold carrier and an additional electron-hole pair. In the case of nMOS transistors, additional electrons are collected by the channel and additional holes are evacuated by the substrate.
2. The carrier hit a Si-H bond and break the bond. An interface state is created and the Hydrogen atom is released in the substrate.

The probability to hit either an atom or a Si-H bond is random. And the average energy involved in each process is the same in both case.

This is the reason why the substrate current is monitored during HCI stress. A high substrate current means a large number of created electron-hole pairs and thus an efficient Si-H bond breakage mechanism.

When interface states are created, the threshold voltage is modified and the subthreshold slope is degraded. This leads to lower current, and degrades the operating frequency of integrated circuit.

## Scaling and HCI

Advances in semiconductor manufacturing techniques and ever increasing demand for faster and more complex integrated circuits (ICs) have driven the associated Metal–Oxide–Semiconductor field-effect transistor (MOSFET) to scale to smaller dimensions.

However, it has not been possible to scale the supply voltage used to operate these ICs proportionately due to factors such as compatibility with previous generation circuits, noise margin, power and delay requirements, and non-scaling of threshold voltage, subthreshold slope, and parasitic capacitance.

As a result internal electric fields increase in aggressively scaled MOSFETs, which comes with the additional benefit of increased carrier velocities (up to velocity saturation), and hence increased switching speed, but also presents a major reliability problem for the long term operation of these devices, as high fields induce hot carrier injection which affects device reliability.

Large electric fields in MOSFETs imply the presence of high-energy carriers, referred to as "**hot carriers**". These hot carriers that have sufficiently high energies and momenta to allow them to be injected from the semiconductor into the surrounding dielectric films

such as the gate and sidewall oxides as well as the buried oxide in the case of silicon on insulator (SOI) MOSFETs.

## CMOS Reliability Impact of HCI

The presence of such mobile carriers in the oxides triggers numerous physical damage processes that can drastically change the device characteristics over prolonged periods. The accumulation of damage can eventually cause the circuit to fail as key parameters such as threshold voltage shift due to such damage. The accumulation of damage resulting degradation in device behavior due to hot carrier injection is called "**hot carrier degradation**".

The useful life-time of circuits and integrated circuits based on such a MOS device are thus affected by the life-time of the MOS device itself. To assure that integrated circuits manufactured with minimal geometry devices will not have their useful life impaired, the life-time of the component MOS devices must have their HCI degradation well understood. Failure to accurately characterize HCI life-time effects can ultimately affect business costs such as warranty and support costs and impact marketing and sales promises for a foundry or IC manufacturer.

## Relationship to Radiation Effects

Hot carrier degradation is fundamentally same as the ionization radiation effect known as the total dose damage to semiconductors, as experienced in space systems due to solar proton, electron, X-ray and gamma ray exposure.

## HCI and NOR Flash Memory Cells

HCI is the basis of operation for a number of non-volatile memory technologies such as Electrically Erasable Programmable Read-Only Memory (EEPROM) cells. As soon as the potential detrimental influence of HC injection on the circuit reliability was recognized, several fabrication strategies were devised to reduce it without compromising the circuit performance.

NOR flash memory exploits the principle of hot carriers injection by deliberately injecting carriers across the gate oxide to charge the floating gate. This charge alters the MOS transistor threshold voltage to represent a logic '0' state. An uncharged floating gate represents a '1' state. Erasing the NOR Flash memory cell removes stored charge through the process of Fowler–Nordheim tunneling.

Because of the damage to the oxide caused by normal NOR Flash operation, HCI damage is one of the factors that cause the number of write-erase cycles to be limited. Because the ability to hold charge and the formation of damage traps in the oxide affects the ability to have distinct '1' and '0' charge states, HCI damage results in the closing of the non-volatile memory logic margin window over time. The number of write-erase cycles

at which '1' and '0' can no longer be distinguished defines the endurance of a non-volatile memory.

# Latchup

**Latchup** is a term used in the realm of integrated circuits (ICs) to describe a particular type of short circuit which can occur in an improperly designed circuit. More specifically it is the inadvertent creation of a low-impedance path between the power supply rails of a MOSFET circuit, triggering a parasitic structure which disrupts proper functioning of the part and possibly even leading to its destruction due to overcurrent. A power cycle is required to correct this situation.

The parasitic structure is usually equivalent to a thyristor (or SCR), a PNPN structure which acts as a PNP and an NPN transistor stacked next to each other. During a latchup when one of the transistors is conducting, the other one begins conducting too. They both keep each other in saturation for as long as the structure is forward-biased and some current flows through it - which usually means until a power-down. The SCR parasitic structure is formed as a part of the totem-pole PMOS and NMOS transistor pair on the output drivers of the gates.

The latchup does not have to happen between the power rails; it can happen at any place where the required parasitic structure exists. A spike of positive or negative voltage on an input or output pin of a digital chip, exceeding the rail voltage by more than a diode drop, is a common cause of latchup. Another cause is the supply voltage exceeding the absolute maximum rating, often from a transient spike in the power supply, leading to a breakdown of some internal junction. This frequently happens in circuits which use multiple supply voltages that do not come up in the proper order after a power-up, leading to voltages on data lines exceeding the input rating of parts that have not yet reached a nominal supply voltage.

Yet another common cause of latchups is ionizing radiation which makes this a significant issue in electronic products designed for space (or very high-altitude) applications.

## *Latchup prevention*

It is possible to design chips that are latchup-resistant, where a layer of insulating oxide (called a *trench*) surrounds both the NMOS and the PMOS transistors. This breaks the parasitic SCR structure between these transistors. Such parts are important in the cases where the proper sequencing of power and signals cannot be guaranteed (e.g., in hot swap devices).

Devices fabricated in lightly doped epitaxial layers grown on heavily doped substrates are also less susceptible to latchup. The heavily doped layer acts as a current sink where excess minority carriers can quickly recombine.

Another possibility for a latchup prevention is the *Latchup Protection Technology* circuit. When a latchup is detected, the LPT circuit shuts down the chip and holds it powered-down for a preset time.

Most silicon-on-insulator devices are inherently latchup-resistant. Latchup is the low resistance connection between tub and power supply rails.

## Chapter-6

# Radiation Hardening

**Radiation hardening** is a method of designing and testing electronic components and systems to make them resistant to damage or malfunctions caused by ionizing radiation (particle radiation and high-energy electromagnetic radiation), such as would be encountered in outer space, high-altitude flight, around nuclear reactors, particle accelerators, or during nuclear accidents or nuclear warfare.

Most radiation-hardened chips are based on their commercial equivalents, with some manufacturing and design variations that reduce the susceptibility to interference from electromagnetic radiation. Due to the extensive development and testing required to produce a radiation-tolerant design of a microelectronic chip, radiation-hardened chips tend to lag behind the cutting-edge of developments.

## *Problems caused by radiation*

Environments with high levels of ionizing radiation create special design challenges. A single charged particle can knock thousands of electrons loose, causing electronic noise and signal spikes. In the case of digital circuits, this can cause results which are inaccurate or unintelligible. This is a particularly serious problem in the design of artificial satellites, spacecraft, military aircraft, nuclear power stations, and nuclear weapons. In order to ensure the proper operation of such systems, manufacturers of integrated circuits and sensors intended for the (military) aerospace markets employ various methods of radiation hardening. The resulting systems are said to be **rad(iation)-hardened**, **rad-hard**, or (within context) **hardened**.

## *Major radiation damage sources*

Typical sources of exposure of electronics to ionizing radiation are the Van Allen radiation belts for satellites, nuclear reactors in power plants for sensors and control circuits, residual radiation from isotopes in chip packaging materials, cosmic radiation for spacecraft and high-altitude aircraft, and nuclear explosions for potentially all military and civilian electronics.

- Cosmic rays come from all directions and consist of approx. 85% protons, 14% alpha particles, and 1% heavy ions, together with x-ray and gamma-ray radiation. Most effects are caused by particles with energies between $10^8$ and $2*10^{10}$ eV. The atmosphere filters most of these, so they are primarily a concern for spacecraft and high-altitude aircraft.

- Solar particle events come from the direction of the sun and consist of a large flux of high-energy (several GeV) protons and heavy ions, again accompanied by x-ray radiation.

- Van Allen radiation belts contain electrons (up to about 10 MeV) and protons (up to 100s MeV) trapped in the geomagnetic field. The particle flux in the regions farther from the Earth can vary wildly depending on the actual conditions of the sun and the magnetosphere. Due to their position they pose a concern for satellites.

- Secondary particles result from interaction of other kinds of radiation with structures around the electronic devices.

- Nuclear reactors produce gamma radiation and neutron radiation which can affect sensor and control circuits in nuclear power plants.

- Nuclear explosions produce a short and extremely intense surge through a wide spectrum of electromagnetic radiation, an electromagnetic pulse (EMP), neutron radiation, and a flux of both primary and secondary charged particles. In case of a nuclear war they pose a potential concern for all civilian and military electronics.

- Chip packaging materials were an insidious source of radiation that was found to be causing soft errors in new DRAM chips in the 1970s. Traces of radioactive elements in the packaging of the chips were producing alpha particles, which were then occasionally discharging some of the capacitors used to store the DRAM data bits. These effects have been reduced today by using purer packaging materials, and employing error-correcting codes to detect and often correct DRAM errors.

## *Radiation effects on electronics*

### Fundamental mechanisms

Two fundamental damage mechanisms take place:

- *Lattice displacement*, caused by neutrons, protons, alpha particles, heavy ions, and very high energy gamma photons. They change the arrangement of the atoms in the crystal lattice, creating lasting damage, and increasing the number of recombination centers, depleting the minority carriers and worsening the analog properties of the affected semiconductor junctions. Counterintuitively, higher

doses over short time cause partial annealing ("healing") of the damaged lattice, leading to a lower degree of damage than with the same doses delivered in low intensity over a long time. This type of damage is especially important for bipolar transistors, which are dependent on minority carriers in their base regions; increased losses caused by recombination cause loss of the transistor gain.

- *Ionization effects* are caused by charged particles, including the ones with energy too low to cause lattice effects. The ionization effects are usually transient, creating glitches and soft errors, but can lead to destruction of the device if they trigger other damage mechanisms, e.g. a latchup. Photocurrent caused by ultraviolet and x-ray radiation may belong to this category as well. Gradual accumulation of holes in the oxide layer in MOSFET transistors leads to worsening of their performance, up to device failure when the dose is high enough.

The effects can vary wildly depending on all the parameters - the type of radiation, total dose and the radiation flux, combination of types of radiation, and even the kind of the device load (operating frequency, operating voltage, actual state of the transistor during the instant it is struck by the particle), which makes thorough testing difficult, time consuming, and requiring a lot of test samples.

## Resultant effects

The "end-user" effects can be characterized in several groups:

- **Neutron effects**: A neutron interacting with the semiconductor lattice will displace its atoms. This leads to an increase in the count of recombination centers and deep-level defects, reducing the lifetime of minority carriers, thus affecting bipolar devices more than CMOS ones. Bipolar devices on silicon tend to show changes in electrical parameters at levels of $10^{10}$ to $10^{11}$ neutrons/cm², CMOS devices aren't affected until $10^{15}$ neutrons/cm². The sensitivity of the devices may increase together with increasing level of integration and decreasing size of individual structures. There is also the risk of induced radioactivity caused by neutron activation, which is a major source of noise in high energy astrophysics instruments. Induced radiation, together with residual radiation from impurities in used materials, can cause all sorts of single-event problems during the device's lifetime. GaAs LEDs, common in optocouplers, are very sensitive to neutrons. The lattice damage influences the frequency of crystal oscillators. Kinetic energy effects (namely lattice displacement) of charged particles belong here too.

- **Total ionizing dose effects**: The cumulative damage of the semiconductor lattice (*lattice displacement* damage) caused by ionizing radiation over the exposition time. It is measured in rads and causes slow gradual degradation of the device's performance; total dose greater than 5000 rads delivered to silicon-based devices in seconds to minutes will cause long-term degradation. In CMOS devices, the radiation creates electron–hole pairs in the gate insulation layers, which cause

photocurrents during their recombination, and the holes trapped in the lattice defects in the insulator create a persistent gate biasing and influence the transistors' threshold voltage, making the N-type MOSFET transistors easier and the P-type ones more difficult to switch on. The accumulated charge can be high enough to keep the transistors permanently open (or closed), leading to device failure. Some self-healing takes place over time, but this effect is not too significant. This effect is the same as hot carrier degradation in high-integration high-speed electronics. Crystal oscillators are somewhat sensitive to radiation dose, which alters their frequency; the sensitivity can be greatly reduced by using swept quartz. Natural quartz crystals are especially sensitive.

- **Transient dose effects**: The short-time high-intensity pulse of radiation, typically occurring during a nuclear explosion. The high radiation flux creates photocurrents in the entire body of the semiconductor, causing transistors to randomly open, changing logical states of flip-flops and memory cells. Permanent damage may occur if the duration of the pulse is too long, or if the pulse causes junction damage or causes a latchup. Latchups are commonly caused by the x-rays and gamma radiation flash of a nuclear explosion. Crystal oscillators may stop oscillating for the duration of the flash due to prompt photoconductivity induced in quartz.

- **Systems-generated EMP effects (SGEMP)** are caused by the radiation flash traveling through the equipment and causing local ionization and electric currents in the material of the chips, circuit boards, cables and cases.

- **Single-event effects (SEE)** are phenomena affecting mostly digital devices.

## Digital damage: SEE

Single-event effects (SEE), mostly affecting only digital devices, were not studied extensively until relatively recently. When a high-energy particle travels through a semiconductor, it leaves an ionized track behind. This ionization may cause a highly localized effect similar to the transient dose one - a benign glitch in output, a less benign bit flip in memory or a register, or, especially in high-power transistors, a destructive latchup and burnout. Single event effects have importance for electronics in satellites, aircraft, and other both civilian and military aerospace applications. Sometimes in circuits not involving latches it is helpful to introduce RC time constant circuits, slowing down the circuit's reaction time beyond the duration of an SEE.

- **Single-event upsets (SEU)**, or **transient radiation effects in electronics**, are state changes of memory or register bits caused by a single ion interacting with the chip. They do not cause lasting damage to the device, but may cause lasting problems to a system which cannot recover from such an error. In very sensitive devices, a single ion can cause a multiple-bit upset (MBU) in several adjacent memory cells. SEUs can become **Single-event Functional Interrupts** (**SEFI**) when they upset control circuits, such as state machines, placing the device into

an undefined state, a test mode, or a halt, which would then need a reset or a power cycle to recover.

- **Single-event latchup (SEL)** can occur in any chip with a parasitic PNPN structure. A heavy ion or a high-energy proton passing through one of the two inner-transistor junctions can turn on the thyristor-like structure, which then stays "shorted" (an effect known as latchup) until the device is power-cycled. As the effect can happen between the power source and substrate, destructively high current can be involved and the part may fail. Bulk CMOS devices are most susceptible.

- **Single-event transient (SET)** happens when the charge collected from an ionization event discharges in the form of a spurious signal traveling through the circuit. This is de facto the effect of an electrostatic discharge.

- **Single-event snapback**, similar to SEL but not requiring the PNPN structure, can be induced in N-channel MOS transistors switching large currents, when an ion hits near the drain junction and causes avalanche multiplication of the charge carriers. The transistor then opens and stays opened.

- **Single-event induced burnout (SEB)** may occur in power MOSFETs when the substrate right under the source region gets forward-biased and the drain-source voltage is higher than the breakdown voltage of the parasitic structures. The resulting high current and local overheating then may destroy the device.

- **Single-event gate rupture (SEGR)** was observed in power MOSFETs when a heavy ion hits the gate region while a high voltage is applied to the gate. A local breakdown then happens in the insulating layer of silicon dioxide, causing local overheat and destruction (looking like a microscopic explosion) of the gate region. It can occur even in EEPROM cells during write or erase, when the cells are subjected to a comparatively high voltage.

## SEE Testing

While proton beams are widely used for SEE testing due to availability, at lower energies proton irradiation can often underestimate SEE susceptibility. Furthermore, proton beams expose devices to risk of total ionizing dose (TID) failure which can cloud proton testing results or result in pre-mature device failure. White neutron beams—while ostensibly the most representative SEE test method—are usually derived from solid target-based sources, resulting in flux non-uniformity and small beam areas. White neutron beams also have some measure of uncertainty in their energy spectrum, often with high thermal neutron content.

The disadvantages of both proton and spallation neutron sources can be avoided by using mono-energetic 14 MeV neutrons for SEE testing. A potential concern is that mono-energetic neutron-induced single event effects will not accurately represent the real-world

effects of broad-spectrum atmospheric neutrons. However, recent studies have indicated that, to the contrary, mono-energetic neutrons—particularly 14 MeV neutrons—can be used to quite accurately understand SEE cross-sections in modern microelectronics.

A particular study of interest, performed in 2010 by Normand and Dominik, powerfully demonstrates the effectiveness of 14 MeV neutrons.

The first devoted SEE testing Laboratory in Canada is currently being established in Southern Ontario under the name Radiation Effects Laboratories.

## *Radiation-hardening techniques*

- Physical:
  - Hardened chips are often manufactured on insulating substrates instead of the usual semiconductor wafers. Silicon oxide (SOI) and sapphire (SOS) are commonly used. While normal commercial-grade chips can withstand between 50 and 100 gray(5 and 10 krad), space-grade SOI and SOS chips can survive doses many orders of magnitude greater. At one time many 4000 series chips were available in radiation-hardened versions (RadHard).
  - Bipolar integrated circuits generally have higher radiation tolerance than CMOS circuits. The low-power Schottky (LS) 5400 series can withstand 1000 krad, and many ECL devices can withstand 10 000 krad
  - Magnetoresistive RAM, or MRAM, is considered a likely candidate to provide radiation hardened, rewritable, non-volatile conductor memory. Physical principles and early tests suggest that MRAM is not susceptible to ionization induced data loss.
  - Shielding the package against radioactivity, to reduce exposure of the bare device.
  - Capacitor-based DRAM is often replaced by more rugged (but larger, and more expensive) SRAM.
  - Choice of substrate with wide band gap, which gives it higher tolerance to deep-level defects; e.g. silicon carbide or gallium nitride.
  - Shielding the chips themselves by use of depleted boron (consisting only of isotope Boron-11) in the borophosphosilicate glass passivation layer protecting the chips, as boron-10 readily captures neutrons and undergoes alpha decay.
- Logical:
  - Error correcting memory uses additional parity bits to check for and possibly correct corrupted data. Since radiation effects damage the memory content even when the system is not accessing the RAM, a "scrubber" circuit must continuously sweep the RAM; reading out the data, checking the parity for data errors, then writing back any corrections to the RAM.
  - Redundant elements can be used at the system level. Three separate microprocessor boards may independently compute an answer to a

calculation and compare their answers. Any system that produces a minority result will recalculate. Logic may be added such that if repeated errors occur from the same system, that board is shut down.

- Redundant elements may be used at the circuit level. A single bit may be replaced with three bits and separate "voting logic" for each bit to continuously determine its result. This increases area of a chip design by a factor of 5, so must be reserved for smaller designs. But it has the secondary advantage of also being "fail-safe" in real time. In the event of a single-bit failure (which may be unrelated to radiation), the voting logic will continue to produce the correct result without resorting to a watchdog timer. System level voting between three separate processor systems will generally need to use some circuit-level voting logic to perform the votes between the three processor systems.
- A watchdog timer will perform a hard reset of a system unless some sequence is performed that generally indicates the system is alive, such as a write operation from an onboard processor. During normal operation, software schedules a write to the watchdog timer at regular intervals to prevent the timer from running out. If radiation causes the processor to operate incorrectly, it is unlikely the software will work correctly enough to clear the watchdog timer. The watchdog eventually times out and forces a hard reset to the system. This is considered a last resort to other methods of radiation hardening.

## Chapter-7

# Failure Modes of Electronics

Electronic devices have a wide range of failure modes. These can be distinguished by their development in time (sudden failure or gradual degradation), by environmental effects (e.g. corrosion, ionizing radiation) or by the electrical parameter which was exceeded (e.g. electrostatic discharge, overvoltage, overcurrent, etc.).

Failures most commonly occur at the beginning and near the end of the lifetime of the parts. Burn-in procedures are used to detect early failures.

Presence of parasitic structures, irrelevant for normal operation, may become important in the context of failures; such structures can be both a source of failure and a protective device.

A sudden fail-open failure can cause multiple secondary failures, when the event is fast and the circuit contains an inductance. The suddenly interrupted current flow in combination with the inductance then causes large voltage spikes, which for very fast events may exceed 500 volts. A burned metallization on a chip may then cause secondary overvoltage damage.

### *Packaging-related failures*

Electronic packaging, acting as the barrier between the materials of the electronic parts and the environment, is very susceptible to environmental factors. Thermal cycling may cause material fatigue due to mechanical stresses induced by thermal expansion, especially when the thermal expansion coefficients of the materials are different. Humidity or presence of aggressive chemicals can cause corrosion of the packaging materials, leads, or cause failure of encapsulation and following damage to the part inside, leading to electrical failure. Exceeding the allowed environmental temperature, whether too high or too low, can cause overstressing of the wire bonds inside the package, tearing the connections, cracking the semiconductor dies, or causing cracks to the packaging itself. Absorption of humidity into the packaging material and subsequent heating to high temperature (e.g. during soldering) may also cause cracking. Mechanical damage could also fall here.

Majority of failure of electronics parts is typically packaging-related.

Bonding wires can be severed or shorted together during encapsulation, or touch the chip die, usually its edge. Dies can crack due to mechanical overstress or thermal shock; initial defects introduced during e.g. wafer sawing or scribing can develop to fractures later. Lead frame may contain excessive material or burrs, causing shorts. Ions of e.g. alkali metals or halogens can be released from the packaging materials and migrate to the semiconductor dies, causing corrosion or parameter deterioration.

Glass-metal seals commonly fail by forming radial cracks. The cracks originate at the pin-glass interface and continue outwards; other failure causes are weak oxide layer on the pin-glass interface and poor formation of glass meniscus around the pin.

Moisture and other gases may be present in the package cavity, either as impurities trapped during manufacture, or from outgassing of the materials used (monomers, curing agents, etc.), or even from chemical reactions (e.g. when the packaging material gets overheated; the released reaction products, often of ionic nature, then can facilitate corrosion and cause a delayed failure). Helium is often added into the inert atmosphere of the packagings, as a tracer gas to detect leaks during testing. Carbon dioxide may result from oxidation of organic materials with residual oxygen. Hydrogen may be released from some organic materials. Moisture can be outgassed by polymers. Amine-cured epoxies may outgas ammonia.

Formation of cracks and growth of intermetallics in die attachment materials may lead to formation of voids and delamination, impairing the heat transfer from the chip die to the substrate and heatsink, and cause a thermal-related failure.

As some semiconductors, notably silicon and gallium arsenide, are transparent in infrared, infrared microscopy can be used to check the integrity of die bonding and under-die structures.

Red phosphorus, used as a charring-promoter flame retardant, facilitates silver migration when present in device packaging materials. The phosphorus particles are normally coated with aluminium hydroxide. If this coating is incomplete, the phosphorus particles oxidize to phosphorus pentoxide, which is strongly hygroscopic and reacts with moisture to phosphoric acid. Phosphoric acid acts as a corrosive electrolyte, which together with electric fields facilitates dissolution and migration of silver, forming shorts between adjanced packaging pins, lead frame leads, tie bars and chip mount structures, and/or chip pads. The silver bridge may be interrupted with thermal expansion of the package; disappearance of the failure when the chip is heated with a heat gun and its reappearance after several minutes later is an indication of this problem.

Delamination and thermal expansion may move the chip die relative to the packaging, deforming and possibly shorting and/or cracking the bonding wires.

## *Electrical contacts*

- Failures of soldered joints
- Failures of electrical contacts – mechanical faults, corrosion
- Failures of cables – fraying, breaking of the conductors, corrosion, fire damage

Soldered joints, whether on boards, cables, or inside the electronic parts themselves, can fail in many ways; by electromigration, mechanical overstress, formation of brittle intermetallic layers, or material fatigue due to excessive thermal cycling. Such failures can be apparent only at high or low joint temperatures, hindering the debugging.

Thermal expansion mismatch between the part package and the printed circuit board material stresses the part-to-board bonds. Leaded parts are able to absorb the strain by bending. Leadless chip packages rely on the properties of the solder to absorb the stresses. Thermal cycling may lead to fatigue cracking of the solder joints, especially with less plastic solders. Various approaches are used to alleviate the temperature-induced strains.

Loose particles can form in the device cavity; a piece of bonding wire, a fragment of the chip die, flakes of plating, particles of die attachment material, fragments of the case, weld flash, and other materials may migrate inside the packaging cavity and cause shorts, often intermittent and sensitive to mechanical shocks.

Corrosion may cause buildup of oxides and other nonconductive products on the electrical contact surfaces. The contacts, when closed, then show unacceptably high resistance. Corrosion products may migrate and cause shorts.

Tin whiskers can form on the tin-coated metals, e.g. on the internal side of the packagings. Loose whiskers then can cause intermittent short circuits inside the packaging.

## *Printed circuit boards*

Printed circuit boards are vulnerable to environmental influences. The traces are prone to corrosion, the vias can be insufficiently plated-through or insufficiently filled with solder. The traces may be improperly etched, either etched-through entirely, weakened, or insufficiently etched and leaving shorts between traces. The traces may crack under mechanical loads; the thin crack then often causes unreliable circuitboard operation, dependent on the physical warping on the board. Residues of solder flux may facilitate corrosion.

Residues of other materials on the surface of the boards can cause leaks. Polar nonionic compounds can attract water molecules from the atmosphere, forming a thin layer of conductive moist coating between the traces (some antistatic agents act the same way). Ionic compounds, especially chlorides, tend to facilitate corrosion. Alkali metal ions may migrate through plastic packaging and influence the operations of the semiconductors.

Chlorinated hydrocarbon residues may hydrolyze, with release of corrosive chlorides; chlorinated solvent residues trapped in the packagings may cause problems years later. Polar molecules may dissipate high-frequency energy, causing dielectric losses.

Above the glass transition temperature of the boards, the resin matrix softens and becomes significantly susceptible to diffusion of contaminants. As an example, polyglycols from the flux can enter the board and increase its humidity intake, with corresponding deterioration of dielectric and corrosion properties.

Multilayer substrates, using ceramics instead of fiber-reinforced polymers, suffer from mostly the same problems.

**Conductive anodic filaments** (CAF) may grow within the boards, along the fibers of the composite material. The metal is introduced to the vulnerable surface typically from plating the vias, then migrates in presence of ions, moisture, and electrical potential. Drilling damage and poor glass-resin bonding promotes such failures. The formation of CAF usually begins by bonding failure between the glass fiber and the resin matrix, and a layer of adsorbed moisture then provides a channel through which ions and corrosion products migrate. In presence of chloride ions, the precipitated material is atacamite (copper chloride hydroxide); its semiconductive properties then may lead to increased leaks, deteriorated dielectric strength, and short circuits between the traces. Absorbed glycols from the flux residues aggravate the problem. The difference in thermal expansion of the fibers and the matrix also weakens the bond when the board is subjected to high temperature during soldering; the lead-free solders, which generally require higher soldering temperatures, are expected to increase the incidence of CAF. CAF incidence depends on absorbed humidity; below certain threshold it does not occur.

Delamination can occur, separating the board layers, cracking the vias and conductors and introducing pathways for corrosive contaminants and migration of conductive species.

## *Semiconductors*

- Reliability (semiconductor)

Many failures result in generation of large amount of hot carriers in the chip structure, namely hot electrons. These are observable under an optical microscope, as they generate near-infrared photons detectable by e.g. a CCD camera. Latchups can be observed this way.

The location of the failure site, if visible, on the chip die may present clues to the nature of the overstress; whether the site is located at the place with highest current density, highest temperature, the highest electric field gradient, etc., size of the damage, secondary damage (fused leads, cracked die, reflowed die attachment...).

Liquid crystal coatings can be used for localization of faults. Cholesteric liquid crystals respond to temperature, are thermochromic; these are used for visualisation of locations of heat production on the chips. Nematic liquid crystals respond to voltage; these are used for visualising current leaks through oxide defects, and for visualising of charge states on the chip surface, allowing seeing the logical states on the individual structures and conductors.

During laser marking of plastic-encapsulated packages, the laser beam may reach and damage the chip die, if the glass spheres used as fillers in the epoxy resin packaging material line up in such way that they conduct the laser light to the chip.

## Parameter-related

- GaAs MMICs:
    - Degradation of $I_{DSS}$: caused by gate sinking and hydrogen effects ("hydrogen poisoning"). Most common and easiest to detect. Affected by reduction of the active channel of the transistor (gate sinking) or depletion of the donor density in the active channel (hydrogen poisoning).
    - Degradation in gate leakage current: occurs at accelerated life tests or high operation temperatures; suspected to be caused by surface-state effects.
    - Degradation in pinch-off voltage": common failure mode for GaAs devices, operating at high temperature. Primarily results from semiconductor-metal interactions and degradation of gate metal structures. Can be hindered by suitable barrier metal inhibiting diffusion between gold and GaAs. Also can be caused by presence of hydrogen.
    - Increase in drain-to-source resistance: observed in devices operating at high temperature, caused by metal-semiconductor interactions. Caused by gate sinking and ohmic contact degradation.
    - Degradation in RF performance: caused by multiple factors. Surface-state density and material related effects play major roles.

In some cases, the normal presence of tolerances in the circuits can cause erratic behavior difficult to trace. For example a combination of a weak driver transistor with a higher series resistance together with the capacitance of the gate of the subsequent transistor, each within the normal "good" specifications, can significantly increase the propagation delay of the signal. Such faults can manifest only at very specific environmental conditions, high clock speeds in combination with low (but within specifications) power supply voltages, and/or specific circuit signal states. Significant variations can occur on a single die. Overstress induced damage, by e.g. creating ohmic shunts or lowering transistor output current, can potentially increase such delays, leading to erratic behavior of the circuit. As the propagation delays show significant dependence on power supply voltage, normally allowed fluctuations of power supply voltage can trigger such erratic behavior.

Vias are a common source of unwanted serial resistance on chips. Defective vias show higher resistance than they should have and therefore increase propagation delays. As

their resistivity drops with increasing temperature, degradation of maximum operating frequency of the chip with decreasing temperature is an indicator of such fault.

**Mousebites** are regions of partially missing metalization. The conductor is still present but its width is locally decreased. Such defects usually do not show during electrical testing, but present a major reliability risk. The increased current density in the damaged region may exacerbate electromigration problems. A very significant degree of voiding is needed to create a temperature-sensitive propagation delay.

## Metalization and wire bonding related

Metalization-related faults are more common and more serious cause of FET transistor degradation than bulk semiconductor material processes. Amorphous materials are promising; the lack of grain boundaries hinders interdiffusion and corrosion.

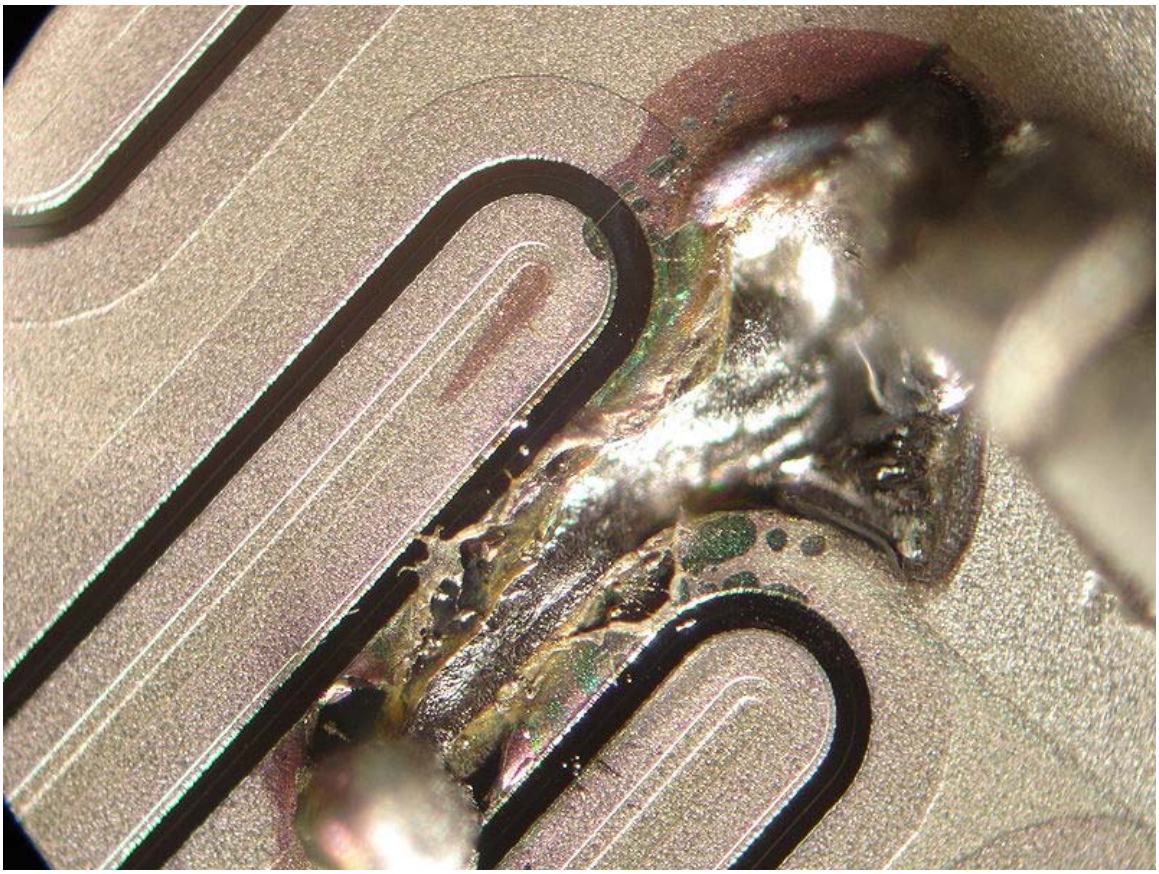
- Electromigration, caused by high current density can move atoms out of the active regions, leading to emergence of dislocations and point defects, acting as nonradiative recombination centers and producing heat instead of light.
    - Al-gate electromigration in power MESFETs can occur with large RF signals. The current densities in the "fingers" of the gate can be sufficient to cause electromigration, leading to voids and interruptions of the gate fingers and consequent loss of control of drain current. Gold metallization is less susceptible, the issue is therefore limited to aluminium.
    - Drain contacts in power FET transistors are depleted on their end while source contacts get material deposited on them.
    - In structures using aluminium metalization over a refractory metal barrier layer, the electromigration affects primarily the aluminium layer. The underlying refractory metal is highly resistant to electromigration, so the conductor does not fail entirely; its resistance just somewhat, often erratically, increases. The displaced aluminium can however cause shorts to neighbouring structures. Addition of 0.5-4% of copper to the aluminium metal significantly increases resistance to electromigration; copper accumulates on the alloy grain boundaries, increasing the energy needed to dislodge the metal atoms from them.

- Metal diffusion caused by high electrical currents or voltages at elevated temperatures can move metal atoms from the electrodes into the active regions. Some materials, notably indium tin oxide and silver, are subject to electromigration which causes leakage current and, in LEDs, nonradiative recombination along the chip edges. A barrier metal layer can be used to hinder the electromigration effects. Metal diffusion can cause changes in dimensions (and therefore parameters) of the transistor gates and other semiconductor junctions. The migration of gate layer in MESFET transistors is known as **gate sinking**; it reduces the dimensions of the active channel and causes change in its effective level of doping, leading to deterioration of electrical parameters.

    - Al/GaAs: Gallium arsenide in contact with aluminium, a common construction of MESFET gates, is susceptible to interdiffusion. Arsenic and especially gallium migrate into the aluminium layer, creating a zone with depleted stoichiometry, and buildup of a $Al_xGa_{1-x}As$ region with gradient of concentration of aluminium. AlAs crystals are formed. This manifests e.g. as an increase of the Schottky barrier height. Forward current accelerates the process in comparison with thermal-only effect. Ti/GaAs shows similar effects. Aluminium-metalized GaAs devices two decades ago had lifetimes of thousands hours; contemporary TiPtAu metalization has lifetimes reaching millions of hours.
    - Au/GaAs: Other interface layers are used with gold; examples are Au/TiW/GaAs, Au/TiPt/GaAs, and Au/Ta/GaAs. However TiW, while best for high temperatures, has different coefficient of thermal expansion than GaAs, which leads to growth of crystal defects under the metalization and reduced carrier mobility. The industry standard structures for gates on GaAs are based on Au/Pt/Ti or Au/Pd/Ti; Ti serves as a thin interlayer to facilitate adhesion, Pt or Pd is a barrier metal hindering diffusion of Au, and Au is the thick layer conductor. Grain boundaries in the barrier metal can however facilitate increased diffusion of gold into GaAs, leading to gate sinking.

- Ohmic contact degradation. The boundary between a metalization layer and the semiconductor can degrade. In case of GaAs, a layer of gold-germanium-nickel alloy (or gold-germanium alloy) is used to achieve low contact resistance. The ohmic contact is achieved by diffusion of germanium into GaAs, forming a highly n-doped region under the metal that facilitates the connection. A thick layer of gold is then deposited over the thin layer of AuGe. Gallium atoms can migrate through the thin layer and get scavenged by the gold above, creating a defect-rich Ga-depleted zone under the contact. Gold (and oxygen) migrate in the other direction, resulting in increased resistance of the ohmic contact and depletion of effective doping level. Formation of intermetallic compounds also plays a role.

- Short circuits; mechanical stresses, high currents, and corrosive environment can lead to formation of whiskers, causing short circuits. These effects can occur both within the packaging of individual devices and on the level of circuit boards.



Micro-photograph of a failed TO3 power transistor due to short circuit

- Formation of intermetallic compounds, e.g. the well-known gold-aluminium intermetallics (the dreaded white and purple plagues), leading to increased contact resistance and vastly decreased mechanical reliability. This is limited to older devices.

- Formation of silicon nodules. Aluminium interconnects may be doped with silicon to saturation during deposition, to prevent alloy spikes. During thermal cycling, the silicon atoms, originally homogeneously distributed, may migrate and clump together, forming nodules. The nodules act as voids in the metallization, increasing its local resistance and lowering the device lifetime.

- Corrosion of metallization. Aluminium is highly susceptible to humidity; especially negatively biased structures can be interrupted easily. Degradation of aluminium metalization was a common cause of failures of early plastic-encapsulated integrated circuits. Gold is susceptible to anodic corrosion in presence of humidity, forming voluminous conductive gold(III) hydroxide. Nickel can be extruded from metalization in presence of humidity and electric field, forming filaments along the electric field gradient that can short the electrodes. Arsenic can be leached from GaAs in presence of moisture. To cause **dry**

**corrosion** of aluminium, only trace amounts of water and ionic contaminants are required; all plastics are somewhat permeable. Phosphate ions can be leached from phosphosilicate glasses used for passivation of the chips; overcoat with oxide or oxynitride layer can be used in modern processes to create a moisture barrier on top of the phosphate glass layer, and/or a borophosphosilicate glass can be used to lower the phosphate content. Halogenides also rapidly corrode aluminium. Chlorides can be transported through the packaging from the outside. Bromides can be liberated from brominated flame retardants present in the packaging plastics when heated above 250 °C; overheating such package during storage, manufacture or use can increase the part's susceptibility to corrosion in the future.

- Sodium contamination, together with less common lithium and potassium. These ions are mobile in silicon dioxide layers even at normal temperature. (Their counterpart anions in contrast stay immobilized in the oxide structure.) The electric fields present during operation of the semiconductors cause migration of the mobile ions, leading to buildup of charged areas in the gate oxide. The NMOS gates are especially susceptible; a positive gate bias repels the cations towards the junction, depressing its threshold voltage. Even low concentrations of mobile ions can cause shifts by few millivolts, enough to cause trouble for analog circuits. The ion migration is a slow process, causing slow gradual drift of circuit parameters up to a possible failure. This effect was observed especially in early metal-gate CMOS logic circuits. Baking the affected chips at 200 °C for a few hours can temporarily reverse the effect by rediffusing the ions through the oxide. Doping the polysilicon of the base region with phosphorus is an effective way for immobilizing the ions. Alkali ions can migrate from the outside of the package; plastics can hinder their movement but can not slow them entirely. Nitride or phosphate glass layers are used as a chip die protection against externally originated contaminants.

- Very narrow interconnection cracks can maintain functionality by electron tunneling across the crack. Cold temperatures cause contraction of the metal, widening the crack and reducing the tunneling. Devices suffering this failure can run at higher frequencies at higher temperatures; this unusual behavior is symptomatic for an interconnection crack.

- Metalization step coverage, or microcracking, is an insidious unscreenable manufacturing failure. The metallization layer at some places, where their height differs, due to geometrical constraints and deposition technique forms locally weakened sites. The metal on such steps is thinner than required, or even develops a microcrack. The increased current density then leads to other effects leading to interruption of the layer and premature failure of the device.

## Semiconductor-related

- Nucleation and growth of dislocations are known mechanisms for degradation of the semiconductor junctions. This requires a presence of an existing defect in the crystal and is accelerated by heat, high current density, and emitted light. In case of LEDs, Gallium arsenide and aluminium gallium arsenide are more susceptible to this mechanism than gallium arsenide phosphide and indium phosphide. Because of different properties of the active regions, gallium nitride and indium gallium nitride are virtually insensitive to this kind of defect.

- Accumulation of charge carriers trapped in the gate oxide of MOSFETs. This introduces permanent gate biasing, influencing the transistor's threshold voltage. This effect may be caused by hot carriers injection, ionizing radiation (one of the total dose effects), or even nominal use; in case of EEPROM cells and related structures this is the major wear mechanism limiting the number of erase-write cycles.

- Ionizing radiation and neutron radiation have multiple effects, both transient and permanent. They can cause defects in the semiconductor, creating recombination centers and shortening the lifetime of minority carriers, degrading the performance of bipolar junction transistors; it also causes accumulation of charge carriers discussed above. More details are described in the problematics of radiation hardening.

- Migration of charge carriers from floating gates, limiting the lifetime of stored data in EEPROM and flash EPROM structures.

- Improper passivation; corrosion effects are a significant source of delayed failures. Semiconductor materials, metallic interconnects, and passivation glasses are all susceptible to corrosion. The surface of semiconductor, subjected to moisture, develops a layer of oxide; the hydrogen liberated from water then reacts with deeper layers of the material, yielding volatile hydrides.

- Stress Induced Leakage Current is an increase in the gate leakage current of a MOSFET, due to defects created in the gate oxide during electrical stressing.

- Hot carrier injection, occurring in MOSFET transistors. At high field gradients, the charge carriers (electrons in NMOS, holes in PMOS) are accelerated to high speeds. When the transistor operates at saturation with high drain-source voltage, the pinched-off region under the gate (created by the electric field) produces hot carriers near the drain end. Due to lower mobility of holes, the NMOS devices produce hot carriers at lower field gradients than PMOS; e.g. a 3 micrometer NMOS will generate hot electrons at 10 V, while a PMOS of the same size will not generate hot holes below 20 V. Some of the hot carriers, after collision with atoms in the drain region, are deflected into the gate oxide. Most return back, but some become trapped in the oxide imperfections. These then accumulate and

cause a persistent gate bias, increasingly shifting the threshold voltage as more hot carriers accumulate. Like with alkali ions, the damage can be partially or fully annealed by heating the chip without power at 200-250 °C for a couple of hours. Avalanche breakdown is also a major source of hot carriers; in case of diffused junction it usually occurs near the surface, where the dopant concentration is highest.

- Parasitic channels can form anywhere where a source-drain structure exists. The role of the gate can be played by a metallization trace above, or even by static charge built up or migrated into the overlaying insulator, protective overcoat, or passivation layer. As the migrating charge carriers are usually electrons, parasitic PMOS transistors tend to be formed. The bipolar chips are more sensitive, likely because of more stringent requirements regarding absence of mobile ionic species for CMOS and BiCMOS technologies.

- Burnt fuses, used for programming the integrated circuits, can under certain conditions reform. Polysilicon fuses may crack instead of vaporizing, which prematurely interrupts the current and leaves enough material to allow reforming the fuse later. Programming both polysilicon and metalization fuses after packaging the circuit prevents the metal from dispersing; the metal stays in the vicinity of the burnt fuse and can migrate back.

- Random access memory chips suffer from several types of failures:
    - Address fault, where the memory address decoder is faulty
    - Stuck-at fault, where a data line somewhere on the chip is shorted to H or L
    - Bridging fault, where two inputs or outputs are shorted together
    - Transition fault
    - n-cell coupling fault
    - Delay fault, where a propagation delay somewhere in the circuit is unacceptably high, causing faulty operation at high speeds
    - Retention fault, where a DRAM cell capacitor does not reliably hold charge for sufficient time

The failures are also classed as bit failures (single, double, triple, quadruple, multiple), row failures (where the entire row of bits fails), column failure (same for the memory array column), cross (where both a row and a column failure occurs), continuous block, peripheral logic, and systemic defect.

## Stress-related

Most stress-related failures are electrothermal in nature. The locally increased temperature can lead to immediate failure by melting or vaporizing metalization layers, melting the semiconductor, or creating other structural changes. The diffusion and electromigration effects tend to be accelerated by high temperature, which shortens the

lifetime of the device. Damages to junctions that do not lead to immediate failure manifest as altered current-voltage characteristics of the junctions.

Electrical overstress failures can be classified as thermally induced failures, electromigration related failures, and electric field related failures.

- Thermal runaway: Nonhomogenities in the substrate, causing localized loss of thermal conductivity, can cause thermal runaway where heat causes damage which causes more heat etc. Most common ones are voids caused by incomplete soldering, or by electromigration effects and Kirkendall voiding.

- Current crowding, non-homogenous distribution of the current density over the junction, formation of current filaments. This may lead to creation of localized hot spots, which poses risk of thermal runaway.

- Reverse bias: Although e.g. the LED is based on a diode junction and is nominally a rectifier, the reverse-breakdown mode for some types can occur at very low voltages and essentially any excess reverse bias causes immediate degradation, and may lead to vastly accelerated failure. 5 V is a typical, "maximum reverse bias voltage" figure for ordinary LEDs, some special types may have lower limits.

- Overcurrent can cause failures of the bonding wires. In some cases the semiconductor junctions can withstand high enough current to melt the bonding wires.

- Zener diodes in reverse bias, when severely overloaded, fail as a short circuit. A sufficiently high voltage causes an avalanche breakdown of the Zener junction; the voltage across the junction together with a significant current being forced through causes extreme localized heating; the junction and metallization melt, and an alloy of silicon and aluminium shorts the diode's terminals. This is sometimes intentionally used in semiconductors as a type of programming fuses.

- Latchups can occur when the device is subjected to an overvoltage or undervoltage pulse. The opened SCR parasitic structure then can cause an overcurrent-based permanent failure. In integrated circuits, latchups are divided by their cause to internal (transmission line reflections, ground bounces, power supply overshoots) and external (signals injected to the chip via its I/O pins from the outside). External latchups can be triggeded by an electrostatic discharge. Effects of ionizing radiation and cosmic rays are also included in external effects. Susceptibility to latchup is tested according to the JEDEC78 test standard. Latchup can be triggered by charge carriers injected into the chip substrate by e.g. a current flowing through an ESD protection diode, or through another latchup.

- FET transistors are sensitive to dV/dt failures; excessively fast voltage transients can cause the transistor to open.

- Bipolar transistors are more thermally sensitive than FETs; the thermal runaway phenomenons limit their operation margins at higher ambient temperatures. Bipolar transistors are also more sensitive to degradation of die cooling mechanisms (die bond defects, part-heatsink attachment degradation).

## Electrostatic discharge

Electrostatic discharge (ESD), a subclass of electrical overstress (EOS), may cause immediate failure of the semiconductor device, a permanent shift of its parameters, or a latent damage causing increased rate of degradation. ESD failure has at least one of three components: localized heat generation, high current density, and high electric field gradient. Currents of several amperes can be present for several hundreds of nanoseconds; the energy deposited to the device structure then causes the damage.

ESD discharge in a real circuit containing capacitances and/or inductances causes a "ringing" waveform, a damped wave with rapidly alternating polarity. Affected junctions are rapidly stressed in alternating forward and reverse polarization.

ESD damage has four basic mechanisms:

- **Oxide rupture/breakdown**, occurring at field strengths above 6–10 MV/cm. Designing the circuit so the oxide layers are protected by junctions with lower avalanche breakdown thresholds than the oxide breakdown voltage can protect against this failure mode.
- Junction damage manifests as increased reverse-bias leakage, up to total shorting.
    - **Junction filamentation**, also known as second or thermal breakdown. Localized overheating causes melting of the silicon; molten silicon has 30 or more times lower resistance; the current through the junction therefore flows in few narrow filaments of high current, resulting in a thermal runaway. In MOSFET devices, the filamentation region is usually near the surface, in the gate-drain overlap region, where the insulator layer serves as a thermal insulation. Devices where the hot spots occur deeper in the structure are less susceptible and are often used as ESD protection structures. Dopants diffuse easily in the molten region; shorter events lead to localized thinning of the base region, resulting in increased leakage, longer events may provide enough time for formation of an ohmic channel across the base, resulting in emitter-collector (or source-drain) short.
    - **Junction spiking**, where metalization is involved; the aluminium metalized devices are more susceptible as the eutectic melting point of aluminium-silicon alloy is 577 °C instead of 1415 °C for silicon. Refractory barrier metals inhibit this effect. When the molten silicon region reaches the metalization, a violent exchange of materials occurs.
- **Metalization and polysilicon burn-out**, where the damage is focused to the conductive and resistive elements – metal and polysilicon interconnects, thin film and diffused resistors. Thin film resistors are the most susceptible. The thermal damage causes localized degradation to destruction of the conductive structure.

The critical current densities can be lower in structures surrounded by thermal insulators where the Joule heat can not easily dissipate.
- **Charge injection**, where the hot carriers generated by an avalanche breakdown are injected into the oxide layer. This mechanism is the same as for hot carrier injection. The result is an increased leakage current from the FETs prebiased by the injected charges in the gate oxide.

For **catastrophic failures**, there are three basic ESD-related mechanisms:

- **Junction burnout**, formation of a conductive path through the junction, shorting it
- **Metalization burnout**, melting or vaporizing part of the metal interconnect, interrupting it

- **Oxide punch-through**, formation of a conductive path through the insulating layer between two conductors or semiconductors; the gate oxides are thinnest and therefore most sensitive. The damaged transistor shows a low-ohmic junction between gate and drain terminals.

ESD can cause a **parametric performance failure**; the device still operates, but its parameters are shifted. The failure may manifest in stress testing. In some cases, the degree of damage can lower over time (so called *cold healing*).

ESD can also cause **latent failures**, which manifest themselves in a delayed fashion and are difficult to impossible to test for. They have these main reasons:

- Damage to insulators: weakening of the insulator structures, leading to accelerated breakdown and/or increased leakage
- Damage to junctions: lowering the lifetime of minority carriers with consequent bipolar transistor gain loss; increasing resistance in forward biased state; increasing leakage in reverse biased state
- Damage to metalization: weakening of the conductor, leading to increased resistance or increased rate of electromigration

Catastrophic failures require the highest discharge voltages; they are easiest to test for, and rarest to occur. Parametric performance failures occur at intermediate discharge voltages and occur more often. Latent failures occur at low voltages and are the most common; for each parametric performance failure there are 4–10 latent ones.

Modern high-integration circuits are more ESD sensitive. The features are smaller, their capacitance is lower and for the same amount of charge the deposited voltage is higher. The silicidation of the conductive layers makes them more conductive, reducing the ballast resistance that can have a degree of protective role.

The gate oxide of some MOSFET transistors can be damaged by as little as 50 volts potential. The gate is isolated from the junction; potential accumulated on it causes

extreme stress on the thin dielectric layer. Stressed oxide can shatter and fail immediately (gate rupture – occurs in nanoseconds, does not require a sustained electric current, and is irreversible; usually the gate and backgate of the affected transistor end up connected together). The gate oxide does not have to fail immediately; the gate leakage can increase by stress induced leakage current, the oxide damage can lead to a delayed failure after hundreds or thousands of operation hours. On-chip capacitors using oxide or nitride dielectric are vulnerable to the same kind of damage. Small structures are more vulnerable than large ones by the virtue of their lower capacitance; the same amount of charge carriers will charge the capacitor the structure forms to a higher voltage. All thin layers of dielectric, e.g. the protective oxide layers over emitter regions of transistors or the insulation between two interconnections, are vulnerable; not just the MOS gates. Chips made by processes employing thicker oxide layers are less vulnerable.

The degree of gate oxide damage depends on the size of the gate. Older transistors had larger gate regions with higher capacitance; a discharge of the accumulated potential frequently caused a gate rupture, causing a **hard breakdown**. Newer ultrathin oxide layers are more commonly damaged by a **soft breakdown**; while the damage is still irreversible, its most significant effect is an increase of the noise voltage of the gate, by up to 4 orders of magnitude. Gate oxide breakdown does not always have to lead to failure; gate-to-channel breakdowns usually do not lead to a hard failure, unlike the gate-to-source or gate-to-drain breakdowns.

Gate oxides can also be stressed during manufacture, by so called **antenna effect**; charges introduced during e.g. ion implantation or dry etching accumulate in conductive structures, causing a voltage buildup across the vulnerable dielectrics.

Breakdown of the gate oxide can form a number of different structures with varying voltage-current characteristics. The simplest case is a connection between two regions with the same doping; a resistive path is then formed. In case of opposite doping regions, the resulting structure is a diode. A metal-to-semiconductor connection may create an ohmic contact or a Schottky diode.

The formation of an ohmic path across an insulator or a junction may not always lead to a hard failure. If the resistance of the path is higher than a critical value, only a degradation of parameters occurs.

Interruption of a conductor may also not always lead to a hard failure. Capacitive coupling can still occur between the structures, maintaining partially degraded functionality, reducing signal strength or affecting gate output voltage range. Faults in drain or source connections of CMOS transistors may lead to formation of quasi-memory cells, making the defect manifestation dependent on previous states of the logic circuit. Loss of an output driver can lead to high-impedance output at some circuit states; the load capacitance then can maintain previous state for some time as a dynamic memory cell.

Semiconductor junctions are more robust than thin dielectrics. A sufficiently high voltage leads to an avalanche breakdown. The electric current is usually concentrated to a small area, due to current crowding; a semiconductor heated above certain limit starts losing resistance with increasing temperature and the region concentrates more current in itself, leading to more heat production and a thermal runaway. The extreme current density can migrate the metalization and short the junction, the heat can melt and recrystallize or even shatter the junction. Such damage usually manifests as a shorted junction.

Current-induced failures are more common in bipolar junction devices, where Schottky or pn junctions are predominant. The high power of the discharge (often above 5 kilowatts for less than a microsecond) is capable of melting and vaporizing silicon and metal. Thin-film resistors can have their value altered by a discharge path forming a shunt across part of their length (value decrease) or get part of the layer vaporized (value increase); this can be problematic in precision analog applications where the values are critical.

A junction which suffered an avalanche breakdown typically has increased leakage. Avalanching the base-emitter junction of a NPN transistor permanently lowers its beta.

Device terminals connected only to MOS gates or capacitors are the most vulnerable to the effects of electrostatic discharge. In case both insulated gates and junctions are connected, avalanching usually takes precedence before gate oxide damage. Substrates of chips and large diffusion areas in the semiconductors, e.g. collectors of high-power transistors, are less vulnerable due to their large size and ability to dissipate more energy. Small diffusion areas, e.g. base and emitor regions of small NPN transistors, are more vulnerable.

Newer CMOS output buffers using a lightly doped drain and silicide source/drain are more ESD sensitive. The N-channel driver usually suffers damage, usually in the oxide layer or n+/p well junction. The damage is caused by current crowding during the snapback of the parasitic NPN transistor. In PMOS-NMOS totem-pole structures, the NMOS transistor is virtually always the one damaged.

Typical ESD-related failure distribution for bipolar/Schottky devices is 90% junction burnout and 10% metalization burnout. For MOSFET devices it is 63% metalization burnout and 27% oxide punchthrough. MIL-HDBK-263 offers more complete discussion.

The structure of the junction influences its ESD sensitivity. Corners and defects can lead to current crowding, reducing the damage threshold. Forward-biased junctions are less sensitive than reverse-biased; in the first case the Joule heat is dissipated through the thicker layer of the material while in the latter it is concentrated in the narrow depletion region.

LEDs and lasers grown on sapphire substrate are more susceptible to ESD damage.

In NMOS transistors, the ESD pulse can lead to formation of a metal filament between source and drain, by a phenomenon called *electrothermomigration*. The junction heat can also melt the polysilicon gate, forming polysilicon filaments between gate-source and gate-drain.

The damage can be observed on the I/V curve; undamaged devices show sharp knees, while damaged ones are significantly softened.

## Optoelectronics

- List of LED failure modes
- Catastrophic optical damage of semiconductor lasers at high power, when the part is overstressed
- Hydrogen darkening of optical fibers in presence of hydrogen
- Phosphor degradation, in white LEDs, cathode ray tubes, plasma displays, fluorescent lights, etc.
- Corrosion of optical materials; silicate glasses are attacked by moisture. Lenses and optical fibers are prone to deterioration by condensed moisture. Phosphate glasses are more susceptible to corrosion than silica-based glasses, different formulations of glasses can have vastly different.

## MEMS

Microelectromechanical systems suffer from specific types of failures.

- Stiction causes the moving parts stick to other surfaces on contact. An external impulse can sometimes release the adhesion and restore functionality. Non-stick coatings, reduction of surface contact area, and increased awareness virtually eliminated the problem in contemporary systems.
- Particle contamination can cause failures by particles lodged between the elements and blocking their movements. Conductive particles may short out circuits, namely electrostatic actuators.
- Wear damages the surfaces; debris, removed from the surfaces during their mutual movements, can be a source of particle contamination.
- Fractures may cause loss of mechanical parts
- Material fatigue may induce cracks in moving structures
- Electrostatic discharge
- Oxide charging, causing electrostatic attraction between parts, may lead to electrostatically mediated stiction
- Dielectric breakdown, causing a short circuit and irreversibly damaging the MEMS structure

## Vacuum tubes

- Vacuum tubes and fluorescent lights are susceptible to degradation of hot cathodes, leading to gradual loss of emission of electrons. The vacuum inside the

tubes can be compromised by outgassing of materials inside, diffusion of gases through the envelope, or envelope failure. Burnout of the hot cathode filaments leads to a sudden failure. Cold cathode devices tend to be more reliable.
- Multipactor effect
- Phosphor degradation

## *Passive elements*

### Resistors

Resistors can fail open (going to infinite resistance), fail short (going to close to zero resistance), or their value can increase or decrease under environmental conditions (e.g. corrosion, material aging) or because of exceeding of performance limits (e.g. overheating).

Thin film resistors are formed from a thin film of a suitable material, e.g. chromium or tantalum nitride.

- Resistor#Failure modes
- The value of certain thin film metal resistors can change when they overheat, due to annealing of their crystalline structure. Extreme overheating may lead to open circuit failure. In integrated circuits, diffused resistors are preferable for applications where high transient currents are to be encountered, e.g. ESD protection, as they are in close contact with the semiconductor substrate which serves as additional heatsinking.
- Mechanical defects from manufacturing can cause intermittent problems or failures. Improperly crimped caps on carbon or metal resistors can become loose and lose contant permanently or intermittently, or the resistor-to-cap interface resistance increase can shift the value of the resistor. Leads to the through-hole resistors are welded to the caps by spark or by butt weld; a faulty weld can become loose. Rough handling can cause such defects or lead to manifestation of latent defects. Thin cracks in the ceramic substrate may cause an open fault, often only annoyingly intermittent.
- Deformation of wire-wound resistors can lead to shorting of adjanced loops of resistive wire, causing partial loss of resistance.
- Operation in oxidizing atmosphere causes oxidation to the outer layer of the resistive wire, reducing its active diameter and increasing its resistance. Operation in reducing atmosphere, e.g. in presence of hydrogen, has the opposite effect.
- Ceramic and carbon resistor cores are prone to cracking under mechanical loads and shocks.
- Under overload, the power dissipated on the resistor can cause heating above the maximum rating, and melting or oxidizing of the resistive element. The protective lacquer or polymer is charred and pyrolyzed, with release of the characteristic smell and possible formation of a conductive path. The resistive element may be interrupted or weakened.

- At high voltages, arcing can occur between parts of the resistor surface, with possible formation of conductive paths or vaporization of the resistive material.
- Damage to insulation layer of the end caps can lead to short between the cap and an underlying circuit board trace.
- Carbon composition resistors, rods of carbon/ceramic composite with embedded leads, may absorb water in humid environments during operation or storage, with resistance changes of as much as 15% up or down.
- Wire-wound resistors are prone to corrosion of the resistive wire. Cap crimping and welds between the wire, caps, and leads are also weak points of this type.
- Metal and carbon film resistors, composed of a thin film of resistive material (typically cut into a spiral) on a (typically) ceramic core, are prone to corrosion and electrolysis of the resistive layer, when the protective layer is penetrated by moisture and contaminants. The electrolytic damage can erode the layer up to an open circuit failure; externally applied voltage is needed for this deterioration mode. Conductive contaminants may form bridges between the loops of the spiral and lead to resistance drop.
- Surface-mount resistors can suffer delamination of their structure where dissimilar materials join, e.g. between the ceramic substrate and the resistive layer, or between the resistive layer and the contact terminations.
- Nichrome thin-film resistors in integrated circuits can be attacked by phosphorus from the passivation glass, which corrodes them and increases their resistance.
- Laser-trimmed resistors may develop instabilities due to the heat damage and microcracks in the resistive layer adjanced to the site of the kerf.
- SMD resistors with silver metallization of termination contacts may suffer open-circuit failure in sulfur-rich environment (e.g. close to sulfur-vulcanized rubber), due to buildup of silver sulfide. Presence of either sulfur dioxide (emitted e.g. by heated rubber) and hydrogen sulfide can be the cause. Conformal coatings do not prevent the failure. The failure is however fairly rare.
- Materials may migrate through the resistor structure and alter resistance or cause shorts. Silver can migrate from the electrodes of thick film chip resistors. Silver atoms were detected as far as 100 micrometers from the electrodes. In addition to lowering the resistance, the silver atoms create nonhomogenities in current distribution and lead to current crowding, instability of characteristics over time, and increase of noise.
- Copper dendrites may grow from copper oxide present in some materials (e.g. from the layer facilitating adhesion of metallization to a ceramic substrate), and bridge the trimming kerf.
- The resistive layer of thick film chip resistors may degrade when subjected to overvoltage or overcurrent.
- The electrodes may crack or corrode.

## Potentiometers and trimmers

Potentiometers and trimmers are three-terminal electromechanical parts, containing a resistive path with a wiper contact with adjustable position. All the failure modes listed for resistors apply to these parts too. Additionally, mechanical wear on the wiper and the

resistive layer, together with corrosion, surface contamination, and mechanical deformations, may lead to intermittent path-wiper resistance changes. These are especially annoying with audio amplifiers. Many types are not perfectly sealed, and contaminants and moisture may readily enter the parts. An especially common contaminant is the solder flux. Mechanical deformations, with impaired wiper-path contact, can occur by warpage of the part housing during soldering or mechanical stress during mounting. Excess stress on leads can cause substrate cracking and open failure when the crack penetrates the resistive path.

## Capacitors

Capacitors are characterized by their capacitance, parasitic resistance in series (equivalent series resistance) and parallel (leakage) – both often frequency-dependent and voltage-dependent, breakdown voltage, and dissipation factor. Structurally, capacitors consist of electrodes separated by the dielectric (sometimes soaked with liquid electrolyte), the leads, and the housing. Deterioration of any of these structures may cause shift of parameters or open or short failure. Short failures and increase of leakage are the most common failure modes of capacitors, followed by open failures.

- Dielectric breakdown occurs due to overvoltage or aging of the dielectric material leading to breakdown voltage falling below the operating voltage; some types of capacitors are "self-healing", the internal arcing vaporizes parts of the electrodes around the failed spot of the dielectric and breaks the contact, others form a conductive pathway through the dielectric, leading to a short or to partial loss of dielectric resistance.
- Electrode materials may migrate across the dielectric, forming conductive paths.
- Leads can be separated from the capacitor by rough handling during storage, assembly or operation; this leads to an open failure. The failure can occur inside the packaging, invisible from the outside but measurable.
- Dissipation factor may increase due to contamination of the capacitor materials, whether from manufacture or by penetration along the leads or through the packaging. Flux and solvent residues are a common source of problems.
- Leakage can result from contaminants forming conductive paths across the capacitor plates or from altering the dielectric parameters. Moisture or solvents can be absorbed by the dielectric if the coating cracks. Cracks in dielectric can also form leakage paths.
- Excessive charging or discharging current may cause partial fusing of the electrodes, leading to open failure or to capacitance shift. Excessive charging or discharging currents may also fuse the leads to the electrodes.
- Partial debonding of leads can cause loss of capacitance in multilayer capacitors.
- Delamination of multilayer capacitors may cause partial capacitance loss and deterioration of other electrical parameters.
- Voids, cracks, pores and other defects may impair the parameters of the dielectric, causing leakages, loss of capacitance, and loss of maximum breakdown voltage.
- Cracks in the dielectric may present pathways for metal migration and arcing.
- Chemical attack on the dielectric may degrade the insulation resistance.

- Thermal shocks, especially during soldering, may cause mechanical defects in the parts, and alter (temporarily or permanently) the properties of the dielectrics.
- Capacitors in low-voltage high-impedance applications generally fail due to lowered dielectric resistance. Capacitors in high voltage and/or low impedance applications tend to fail due to dielectric breakdown.
- Pinholes and voids cause increased leakage in film capacitors. When sufficient current is available, the defect site can be burned away and the capacitor "self-heals". When insufficient current is available, the current can cause migration of electrode material through the defect, gradually decreasing the leakage path resistance.
- Film capacitors are sensitive to excessive ripple voltage; the thin metallization on the dielectric may not sustain high currents, and partially fuse with a loss of capacitance.
- Polymer film dielectrics tend to become brittle with age, especially at increased temperature. Fractures may develop. Moisture may cause degradation of the polymer.
- Corrosion and cracking of the internal leads can cause open failure.
- Overvoltage in a ceramic capacitor may cause an avalanche breakdown in a weak spot of the dielectric, resulting in a thermal runaway. The damaged area may show localized melting and cracking. The energy deposited in the failure point can be high enough to cause the capacitor to explode, create a miniature fireball, and char the encapsulation and possibly the underlying circuitboard.
- Mica capacitors may show intermittent open failures from separation of the electrode plates from the termination at certain temperatures. Reconstituted mica capacitors may fail short due to weaknesses in the dielectric. As mica capacitors are typically used in high voltage applications, the results may be severe.

## Electrolytic capacitors

- Aluminium electrolytic capacitors suffer from gradual increase of leakage and equivalent series resistance, and loss of capacitance due to drying out of the electrolyte. Power dissipation due to high ripple currents and high internal resistance cause rise of the internal temperature of the capacitor, which accelerates the deterioration rate, especially when the internal temperature exceeds maximum design temperature. Such capacitors usually fail short.
- Electrolyte contamination, especially with moisture, can lead to corrosion of the electrodes. The deterioration of their area can lead to capacitance loss, loose particles of conductive material can cause shorts.
- Electrolyte may under certain conditions evolve gas. This leads to increased pressure inside the capacitor housing, in extreme but frequent cases leading to an explosion.
    - Capacitor plague is an extreme kind of premature electrolytic capacitor degradation and failure.
- Tantalum capacitors are susceptible to electrical overstress. Exceeding of the maximum surge voltage may permanently degrade the dielectric and even cause open or short failure.

- The most common failure mode of solid tantalum capacitors is an electrical short. The most common manufactured-related cause of this failure is a defect in the dielectric oxide.
- The failure locations may be visible as discolored dielectric, or as locally melted anode.
- Defects in the tantalum chip anode material, e.g. impurities, are transferred to the dielectric oxide layer formed during the capacitor manufacture. These locations then present weak spots that may fail prematurely. Thermal shock e.g. during soldering may generate new defects.
- The thickness of the oxide layer is limited by the size of the powder from which the tantalum slug is made.
- On voltage transients, tantalum capacitors may show a momentary short or increase of leakage current, called "scintillation". If the current through the failure site is small, the site self-heals; if the current is sufficient to heat the site above 400 °C, thermal runaway occurs and the part shorts.

## Crystals

Crystals for crystal oscillators are thin slabs of a piezoelectric material, typically quartz, with deposited electrodes on the slab surfaces, mounted in a hermetically sealed housing, providing the circuit with oscillations at a stable frequency.

- Cracks of crystal slab can occur at extremely high drive levels (especially with low-frequency crystals), or due to mechanical shocks strong enough to make the crystal impacts the inside of the housing.
- Failure of lead connection can lead to an open circuit failure.
- Outgassing of materials inside the housing, most often improperly cured epoxy, can lead to frequency shifts of the crystal.
- Moisture present in the housing can condense on the crystal at low temperatures, causing significant frequency shifts.
- Ionizing radiation and neutron radiation can cause frequency shifts. Swept quartz, from which the mobile alkali metal ions were removed, is used for rad-hard crystals.
- Dehermetization of the housing, by rough handling or corrosion, may lead to penetration of contaminants and moisture into the housing, with consequent changes in resistance and frequency, and possible lead-to-lead and lead-to-packaging leakage paths.

## Chapter-8

# Semiconductor Device Fabrication

NASA's Glenn Research Center cleanroom.

**Semiconductor device fabrication** is the process used to create the integrated circuits (silicon chips) that are present in everyday electrical and electronic devices. It is a multiple-step sequence of photographic and chemical processing steps during which electronic circuits are gradually created on a wafer made of pure semiconducting material. Silicon is the most commonly used semiconductor material today, along with various compound semiconductors.

The entire manufacturing process, from start to packaged chips ready for shipment, takes six to eight weeks and is performed in highly specialized facilities referred to as fabs.

## *History*

When feature widths were far greater than about 10 micrometres, purity was not the issue that it is today in device manufacturing. As devices became more integrated, cleanrooms became even cleaner. Today, the fabs are pressurized with filtered air to remove even the smallest particles, which could come to rest on the wafers and contribute to defects. The workers in a semiconductor fabrication facility are required to wear cleanroom suits to protect the devices from human contamination.

In an effort to increase profits, semiconductor device manufacturing has spread from Texas and California in the 1960s to the rest of the world, such as Europe, Middle East, and Asia. It is a global business today.

The leading semiconductor manufacturers typically have facilities all over the world. Intel, the world's largest manufacturer, has facilities in Europe and Asia as well as the U.S. Other top manufacturers include STMicroelectronics (Europe), Analog Devices (US), Integrated Device Technology (US), Atmel (US/Europe), Freescale Semiconductor (US), Samsung (Korea), Texas Instruments (US), GlobalFoundries (Germany, Singapore, future New York fab in construction), Toshiba (Japan), NEC Electronics (Japan), Infineon (Europe), Renesas (Japan), Taiwan Semiconductor Manufacturing Company (Taiwan), Fujitsu (Japan/US), NXP Semiconductors (Europe), Micron Technology (US), Hynix (Korea) and SMIC (China).

## *Wafers*

A typical wafer is made out of extremely pure silicon that is grown into mono-crystalline cylindrical ingots (boules) up to 300 mm (slightly less than 12 inches) in diameter using the Czochralski process. These ingots are then sliced into wafers about 0.75 mm thick and polished to obtain a very regular and flat surface.

Once the wafers are prepared, many process steps are necessary to produce the desired semiconductor integrated circuit. In general, the steps can be grouped into two major parts:

- Front-end-of-line (FEOL) processing
- Back-end-of-line (BEOL) processing

## *Processing*

In semiconductor device fabrication, the various processing steps fall into four general categories: deposition, removal, patterning, and modification of electrical properties.

- Deposition is any process that grows, coats, or otherwise transfers a material onto the wafer. Available technologies consist of physical vapor deposition (PVD), chemical vapor deposition (CVD), electrochemical deposition (ECD), molecular beam epitaxy (MBE) and more recently, atomic layer deposition (ALD) among others.
- Removal processes are any that remove material from the wafer either in bulk or selectively and consist primarily of etch processes, either wet etching or dry etching. Chemical-mechanical planarization (CMP) is also a removal process used between levels.
- Patterning covers the series of processes that shape or alter the existing shape of the deposited materials and is generally referred to as lithography. For example, in conventional lithography, the wafer is coated with a chemical called a *photoresist*. The photoresist is exposed by a *stepper*, a machine that focuses, aligns, and

moves the mask, exposing select portions of the wafer to short wavelength light. The unexposed regions are washed away by a developer solution. After etching or other processing, the remaining photoresist is removed by plasma ashing.

- Modification of electrical properties has historically consisted of doping transistor sources and drains originally by diffusion furnaces and later by ion implantation. These doping processes are followed by furnace anneal or in advanced devices, by rapid thermal anneal (RTA) which serve to activate the implanted dopants. Modification of electrical properties now also extends to reduction of dielectric constant in low-k insulating materials via exposure to ultraviolet light in UV processing (UVP).

Many modern chips have up to eleven metal levels produced in over 300 sequenced processing steps.

## Front-end-of-line (FEOL) processing

FEOL processing refers to the formation of the transistors directly in the silicon. The raw wafer is engineered by the growth of an ultrapure, virtually defect-free silicon layer through epitaxy. In the most advanced logic devices, *prior* to the silicon epitaxy step, tricks are performed to improve the performance of the transistors to be built. One method involves introducing a *straining step* wherein a silicon variant such as silicon-germanium (SiGe) is deposited. Once the epitaxial silicon is deposited, the crystal lattice becomes stretched somewhat, resulting in improved electronic mobility. Another method, called *silicon on insulator* technology involves the insertion of an insulating layer between the raw silicon wafer and the thin layer of subsequent silicon epitaxy. This method results in the creation of transistors with reduced parasitic effects.

## Gate oxide and implants

Front-end surface engineering is followed by: growth of the gate dielectric, traditionally silicon dioxide ($SiO_2$), patterning of the gate, patterning of the source and drain regions, and subsequent implantation or diffusion of dopants to obtain the desired complementary electrical properties. In dynamic random access memory (DRAM) devices, storage capacitors are also fabricated at this time, typically stacked above the access transistor (implementing them as trenches etched deep into the silicon surface was a technique developed by the now defunct DRAM manufacturer Qimonda).

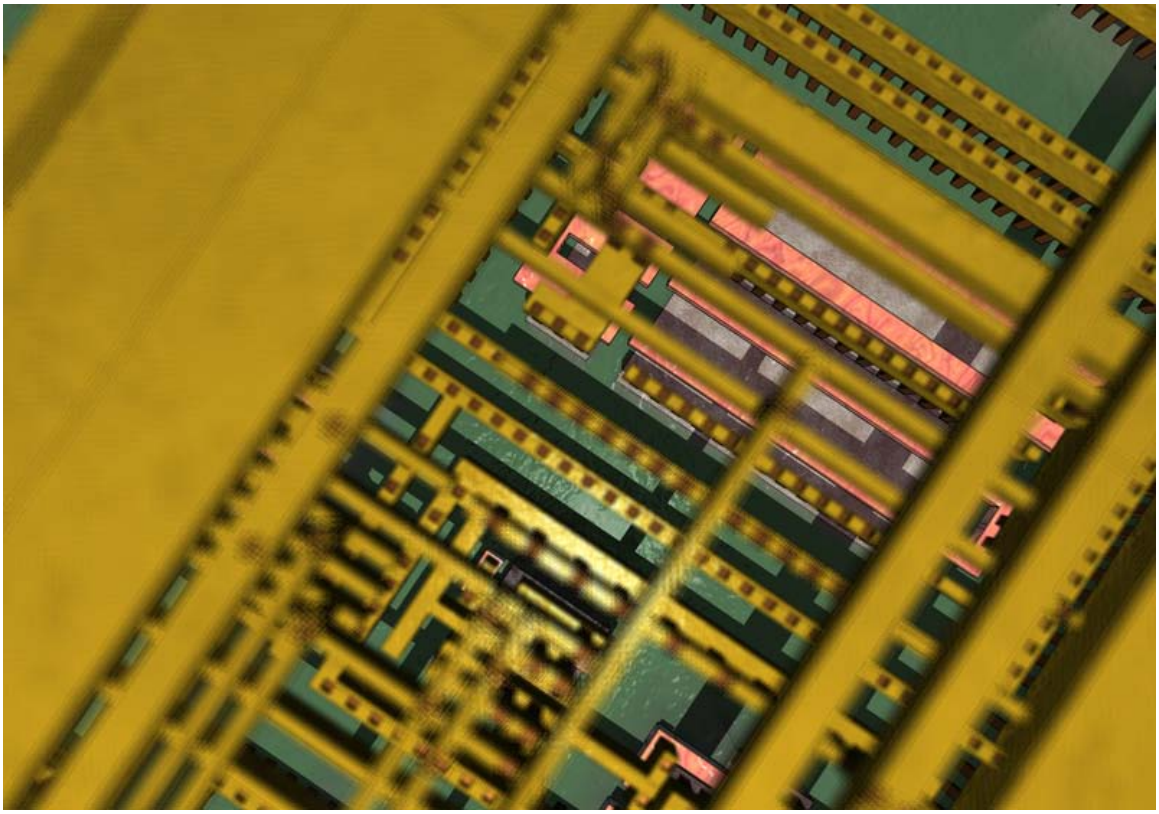
## Back-end-of-line (BEOL) processing

## Metal layers

Once the various semiconductor devices have been created, they must be interconnected to form the desired electrical circuits. This occurs in a series of wafer processing steps collectively referred to as BEOL (not to be confused with *back end* of chip fabrication which refers to the packaging and testing stages). BEOL processing involves creating metal interconnecting wires that are isolated by dielectric layers. The insulating material

was traditionally a form of $SiO_2$ or a silicate glass, but recently new low dielectric constant materials are being used. These dielectrics presently take the form of SiOC and have dielectric constants around 2.7 (compared to 3.9 for $SiO_2$), although materials with constants as low as 2.2 are being offered to chipmakers.

## Interconnect



Synthetic detail of a standard cell through four layers of planarized copper interconnect, down to the polysilicon (pink), wells (greyish) and substrate (green).

Historically, the metal wires consisted of aluminium. In this approach to wiring often called *subtractive aluminium*, blanket films of aluminium are deposited first, patterned, and then etched, leaving isolated wires. Dielectric material is then deposited over the exposed wires. The various metal layers are interconnected by etching holes, called *vias*, in the insulating material and depositing tungsten in them with a CVD technique. This approach is still used in the fabrication of many memory chips such as dynamic random access memory (DRAM) as the number of interconnect levels is small, currently no more than four.

More recently, as the number of interconnect levels for logic has substantially increased due to the large number of transistors that are now interconnected in a modern microprocessor, the timing delay in the wiring has become significant prompting a change in wiring material from aluminium to copper and from the silicon dioxides to

newer low-K material. This performance enhancement also comes at a *reduced cost* via damascene processing that eliminates processing steps. In damascene processing, in contrast to subtractive aluminium technology, the dielectric material is deposited first as a blanket film, and is patterned and etched leaving holes or trenches. In *single damascene* processing, copper is then deposited in the holes or trenches surrounded by a thin barrier film resulting in filled vias or wire *lines* respectively. In *dual damascene* technology, both the trench and via are fabricated before the deposition of copper resulting in formation of both the via and line simultaneously, further reducing the number of processing steps. The thin barrier film, called copper barrier seed (CBS), is necessary to prevent copper diffusion into the dielectric. The ideal barrier film is as thin as possible. As the presence of excessive barrier film competes with the available copper wire cross section, formation of the thinnest continuous barrier represents one of the greatest ongoing challenges in copper processing today.

As the number of interconnect levels increases, planarization of the previous layers is required to ensure a flat surface prior to subsequent lithography. Without it, the levels would become increasingly crooked and extend outside the depth of focus of available lithography, interfering with the ability to pattern. CMP (chemical mechanical planarization) is the primary processing method to achieve such planarization although dry *etch back* is still sometimes employed if the number of interconnect levels is no more than three.

## *Wafer test*

The highly serialized nature of wafer processing has increased the demand for metrology in between the various processing steps. Wafer test metrology equipment is used to verify that the wafers haven't been damaged by previous processing steps up until testing. If the number of dies—the integrated circuits that will eventually become chips— etched on a wafer exceeds a failure threshold (ie. too many failed dies on one wafer), the wafer is scrapped rather than investing in further processing.

## *Device test*

Once the front-end process has been completed, the semiconductor devices are subjected to a variety of electrical tests to determine if they function properly. The proportion of devices on the wafer found to perform properly is referred to as the yield.

The fab tests the chips on the wafer with an electronic tester that presses tiny probes against the chip. The machine marks each bad chip with a drop of dye. Currently, electronic dye marking is possible if wafer test data is logged into a central computer database and chips are "binned" (i.e. sorted into virtual bins) according to predetermined test limits. The resulting binning data can be graphed, or logged, on a wafer map to trace manufacturing defects and mark bad chips. This map can be also used during wafer assembly and packaging. Usually, the fab charges for test time, with prices in the order of cents per second. Test times vary from a few milliseconds to a couple of seconds, and the test software is optimized for reduced test time. multiple chip (Multi-site) testing is also

available, although given the sequential operation of the tester itself (one chip is tested at a time) such arrangement becomes less feasible as test times are increased. Chips are often designed with “testability features” such as "built-in self-test" to speed testing, and reduce test costs. In certain designs that use specialized analog fab processes, wafers are also laser-trimmed during test, to achieve tightly-distributed resistance values as specified by the design.

Good designs try to test and statistically manage *corners*: extremes of silicon behavior caused by operating temperature combined with the extremes of fab processing steps. Most designs cope with more than 64 corners.

## *Die preparation*

Once tested, a wafer is typically reduced in thickness

before the wafer is scored and then broken into individual die -- wafer dicing.

Only the good, unmarked chips go on to be packaged.

## *Packaging*

Plastic or ceramic packaging involves mounting the die, connecting the die pads to the pins on the package, and sealing the die. Tiny wires are used to connect pads to the pins. In the old days, wires were attached by hand, but now purpose-built machines perform the task. Traditionally, the wires to the chips were gold, leading to a “lead frame” (pronounced “leed frame”) of copper, that had been plated with solder, a mixture of tin and lead. Lead is poisonous, so lead-free “lead frames” are now mandated by ROHS.

Chip-scale package (CSP) is another packaging technology. A plastic dual in-line package, like most packages, is many times larger than the actual die hidden inside, whereas CSP chips are nearly the size of the die. CSP can be constructed for each die *before* the wafer is diced.

The packaged chips are retested to ensure that they were not damaged during packaging and that the die-to-pin interconnect operation was performed correctly. A laser etches the chip’s name and numbers on the package.

## *List of steps*

This is a list of processing techniques that are employed numerous times in a modern electronic device and do not necessarily imply a specific order.

- Wafer processing
    - Wet cleans
    - Photolithography

    - Ion implantation (in which dopants are embedded in the wafer creating regions of increased (or decreased) conductivity)
    - Dry etching
    - Wet etching
    - Plasma ashing
    - Thermal treatments
        - Rapid thermal anneal
        - Furnace anneals
        - Thermal oxidation
    - Chemical vapor deposition (CVD)
    - Physical vapor deposition (PVD)
    - Molecular beam epitaxy (MBE)
    - Electrochemical Deposition (ECD).
    - Chemical-mechanical planarization (CMP)
    - Wafer testing (where the electrical performance is verified)
    - Wafer backgrinding (to reduce the thickness of the wafer so the resulting chip can be put into a thin device like a smartcard or PCMCIA card.)
- Die preparation
    - Wafer mounting
    - Die cutting
- IC packaging
    - Die attachment
    - IC Bonding
        - Wire bonding
        - Thermosonic Bonding
        - Flip chip
        - Tab bonding
    - IC encapsulation
        - Baking
        - Plating
        - Lasermarking
        - Trim and form
- IC testing

## *Hazardous materials*

Many toxic materials are used in the fabrication process. These include:

- poisonous elemental dopants such as arsenic, antimony and phosphorus
- poisonous compounds like arsine, phosphine and silane
- highly reactive liquids, such as hydrogen peroxide, fuming nitric acid, sulfuric acid and hydrofluoric acid

It is vital that workers not be directly exposed to these dangerous substances. The high degree of automation common in the IC fabrication industry helps to reduce the risks of exposure of this sort. Most fabrication facilities employ exhaust management systems,

such as wet scrubbers, combustors, heated absorber cartridges etc, to control the risk to workers and also the environment if these toxic materials are released into the atmosphere.